University of Alberta

Emergency Medical Service Systems: Modelling Uncertainty in Response Time

by

Susan D. Budge

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of Doctor of Philosophy

in

Management Science

Faculty of Business

Edmonton, Alberta

Fall 2004

Library and Archives Canada

Published Heritage Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et Archives Canada

Direction du Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence
ISBN: 0-612-95912-0
Our file Notre référence
ISBN: 0-612-95912-0

The author has granted a non-exclusive license allowing the Library and Archives Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque et Archives Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou aturement reproduits sans son autorisation.

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this thesis.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de cette thèse.

While these forms may be included in the document page count, their removal does not represent any loss of content from the thesis.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.

Canada

Acknowledgements

I feel very fortunate to have been surrounded by so many people that have made my experience as a PhD student a rich one that I will remember fondly. It is my pleasure to thank those who contributed in one way or another to my success.

First, and above all, I would like to express my sincere gratitude to my supervisor, Dr. Armann Ingolfsson, for his guidance and support, but especially for his patience, over the course of my program and this research. In addition, I would like to thank my committee members, Dr. Erhan Erkut, Dr. Tarja Joro, and Dr. Ignacio Castillo for all of their help and advice throughout this process. I am also very grateful for the participation and helpful suggestions of my external examiners, Dr. John Hodgson and Dr. Rajan Batta.

Next, I wish to express my appreciation to Dr. David Cooper, Dr. Terry Daniel, Dr. Dave Jobson, Jeanette Gosine, Keltie Tolmie, Louise Hebert, and Kathy Harvey for their assistance whenever it was requested. I was also very lucky to have had this experience with a great group of students in Business PhD program and I would especially like to thank Edgar Cabral, Malgorzata Korolkiewicz, and Yongyue Li who went through the process alongside me and provided encouragement when it was most needed.

I would like to recognize the various organizations that provided the financial support at various stages of my program making this research possible, the Faculty of Business, the Faculty of Graduate Studies and Research, and the Office of Critical Infrastructure Protection and Emergency Preparedness. In addition, I want to express my sincere gratitude to the people from City of Edmonton Emergency Medical Services and the City of Calgary Emergency Medical Services for the generosity that they showed with their time, and for providing the information and data without which this research would not have been possible.

Finally, I want to thank my family. The constant love and support of my mom, my dad, my sister, and my partner, Trevor, have made this goal possible for me.

TABLE OF CONTENTS

Chapter 1: Introduction and Literature Review 1

Introduction 1

Literature Review 4

References 10

Chapter 2: Approximating Ambulance Dispatch Probabilities 15

Motivation and Introduction 15

Literature Review 18

The Approximation Procedure 23

Algorithm 31

Computational Results 32

Convergence 32

Accuracy 35

Impact on Performance Measures 40

Conclusions and Further Research 42

References 43

Chapter 3: Optimal Ambulance Location with Random Delays and Travel Times 46

Introduction 46

Literature Review 53

Problem Data 59

Problem Formulation and Properties 61

Concavity Result 62

Busy Fractions 65

Computational Experiments 66

Discussion 70
Model Extensions 70

Further Research 71

Conclusions 72

References 73

Chapter 4: Empirical Analysis of Ambulance Travel Times 76

Introduction 76

Literature Review 77
Travel Distances 78
Travel Times 80
Randomness in Travel Times 84

Data 87
Format of the Data 87
Data Issues 88
Outlier Analysis 91

Methodology 95

Results 96
Preliminary Analysis 96
Modelling Emergency Vehicle Travel Times 105

Conclusions and Further Research 111

References 113

Chapter 5: Conclusions and Further Research 116

Summary 116

Conclusions 117

Further Research 117

References 119

Appendices 120

Appendix 1: Estimating the Average Busy Fraction 120

Appendix 2: Derivation for the KWH Function 121

Appendix 3: Method for Fitting the KWH Function 123

LIST OF TABLES

Table 2 - 1: Summary of model assumptions of previous literature. 19
Table 2 - 2: Scenarios included in experimental design. 35
Table 2 - 3: Average errors and average relative errors. 36
Table 2 - 4: Results for a particular scenario by station. 37
Table 3 - 1: Six ways to model pre-trip delays and travel times. 47

LIST OF FIGURES

Figure 1 - 1: Sequence of events and time intervals for an EMS response. 2

Figure 2 - 1: Graphs of correction factors for various scenarios. 30

Figure 2 - 2: Time to converge for 55,401 scenarios of Edmonton data. 33

Figure 2 - 3: Relative frequencies of correction factors by station preference. 34

Figure 2 - 4: Graph of absolute error by station. 39

Figure 2 - 5: Estimated busy fractions for a particular scenario. 40

Figure 2 - 6: Comparison of estimated coverage using system-wide busy fractions verses site-specific busy fractions. 42

Figure 3 - 1: Error induced by using constant delay times. 49

Figure 3 - 2: The histogram of pre-trip delays for St. Albert sample. 50

Figure 3 - 3: Distribution of travel times between a particular station demand node pair. 51

Figure 3 - 4: An example of iterating on the busy fractions ρ_j. 67

Figure 3 - 5: Sensitivity of the number of ambulances needed to provide the coverage goal to the parameters of the delay distribution. 68

Figure 3 - 6: Sensitivity of the system wide coverage to the parameters of the delay distribution. 69

Figure 4 - 1: Speed time profile assumed for the KWH function. 81

Figure 4 - 2: Speed time profiles of ambulances responding to events. 90

Figure 4 - 3: Box plots of the travel times for various datasets. 93

Figure 4 - 4: Probability density and cumulative distribution functions for Dataset 6. 94

Figure 4 - 5: Speed time profiles of the travel to scene component of ambulances responding to events. 97

Figure 4 - 6: Box plots for the travel time as a function of distance. 98

Figure 4 - 7: Travel time distribution for a particular demand location. 99

Figure 4 - 8: Cumulative distribution functions and summary box plots for the travel times for a particular demand location, from the closest station and from all other locations. 100

Figure 4 - 9: Cumulative distribution functions and summary box plots for the travel times for a particular demand location, from the closest station grouped by priority. 101

Figure 4 - 10: Cumulative distribution functions and summary box plots for the travel times for a particular demand location, grouped into rush hour and off rush hour responses. ____ **101**

Figure 4 - 11: Empirical and theoretical cumulative distribution functions before and after the removal of outliers. ____ **103**

Figure 4 - 12: Q-Q plot of the log of the ordered travel times vs. the corresponding normal quantiles. ____ **103**

Figure 4 - 13: Empirical and theoretical cumulative distribution functions before and after the removal of outliers. ____ **104**

Figure 4 - 14: Q-Q plot of the log of the ordered travel times vs. the corresponding normal quantiles. ____ **104**

Figure 4 - 15: Scatterplot of travel time vs. distance for original data. ____ **105**

Figure 4 - 16: Scatterplot of travel time vs. distance for original data, focusing on relevant range. ____ **106**

Figure 4 - 17: Scatterplot of the residuals of the estimated KWH travel time function. ____ **107**

Figure 4 - 18: Scatterplot of the median travel times by distance and the estimated KWH travel time function. ____ **108**

Figure 4 - 19: Spread vs. level plot for the travel time divided into level bins based on estimated travel time. ____ **109**

Figure 4 - 20: Scatterplot of the first and second quartiles of the travel times by distance and estimated relationships with the rectilinear distance. ____ **110**

Figure 4 - 21: Estimated relationship of the multiplicative standard deviation of the travel times with the rectilinear distance. ____ **110**

Figure 4 - 22: Pictures of the acceleration, velocity and distance as functions of time for the KWH travel time relation. ____ **121**

Chapter 1: Introduction and Literature Review

Introduction

Time is of the essence in any emergency medical service (EMS) system. One manifestation of this is that almost all performance measures used for such systems have some aspect of time at their foundation. The *response time* (the time from when a call to 911 is placed until paramedics reach the scene) is one of the most critical of these performance measures. Response time is important because one would expect that the sooner an ambulance responds to an emergency call, the better the patient's chances for survival and several studies have confirmed this (for example, Cretin and Willemain, 1979; Eisenberg, Bergner, and Hallstrom, 1979; and Stiell et al., 1999). Geography is also central to the modelling of EMS operations. Since the demand is spread out geographically in these systems, the ability to respond quickly to different locations is a necessity. The sequence of activities and corresponding time intervals for a typical EMS event are shown in Figure 1-1. Note that intervals prior to the receipt of the call are not shown, and many of the intervals shown can be divided further (see for example Spaite et al., 1993), however for the purposes of this dissertation, the intervals shown are sufficient. As indicated in the figure, the response time is made up of the travel (to scene) and pre-travel delay intervals. The rest of the components that make up the unit service time (the on scene, transport, and hospital intervals) can also affect the response time indirectly, through their impact on the availability of the unit. Since units can respond to new calls while enroute from a hospital to a station, the return interval is not included in the unit service time.

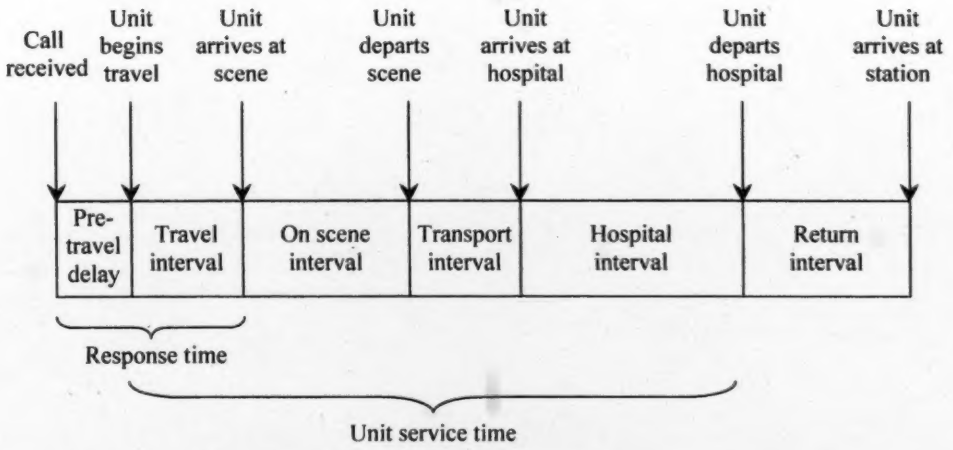


Figure 1-1: Sequence of events and corresponding time intervals for a typical EMS response.

Planning for EMS organizations involves interrelated decisions at many levels, from strategic decisions related to the number and location of stations or to work rules and procedures, to tactical decisions such as staffing and scheduling (allocating paramedic crews and ambulances temporally and spatially), to operational decisions such as real-time repositioning and deployment of units. The research in this dissertation focuses on models that can be used for tactical decisions, although some of the models can be used for certain types of strategic decisions or as building blocks for certain types of operational decisions. Note that planning activities for EMS, fire, and police systems have many things in common, and some (but not all) models developed for use in one of these areas can be, and have been, applied in all three sectors.

A particular experience in modelling the operations of an EMS system (Ingolfsson, Erkut, and Budge, 2003) provided the initial impetus for the work described in this dissertation. That project came about in response to a major operational change under consideration for the EMS system of the City of Edmonton. In a review of the city's EMS services, a recommendation was made

to move to a single start station in which all crews would begin and end their shift at a single location (KPMG and Fitch & Associates, 2000a and 2000b). We developed a simulation model to estimate the impact of such a change on the performance of the system. When considering alternative operational changes for improving the performance, we explored the literature on emergency response vehicle location/allocation models in an effort to narrow down the potential candidates to investigate further within the simulation model. In this investigation, however, none of the models found in the literature was deemed to be completely appropriate for these purposes. In particular, it became apparent that the available prescriptive (optimization) models were based on severely limiting assumptions about the behaviour of the system.

Although there is an extensive literature on models of emergency service operations, including many papers on optimal location of emergency response vehicles and an abundance of case studies, what seems to be lacking is a prescriptive model that is realistic enough to be used to describe the operations of such a system and yet tractable enough to be solved to optimality in a reasonable time. Early papers considered a demand "covered" if there was a vehicle positioned within a certain response time (or distance) standard. As the literature evolved, the focus changed to a more realistic notion of coverage that incorporated the possibility that the vehicle might not be available to respond to an incoming call, in addition to variability in the response time of an available unit. Stochastic performance measures such as the reliability of response within a certain standard time or the fraction of calls responded to within a certain standard time became prevalent. (In this dissertation, the term *coverage* will be used to refer to the fraction of demand that the system can respond to within a specified time.) In order to more accurately calculate such performance measures researchers began to use analytical models, notably the hypercube queueing model (Larson, 1974) and later approximations to this model (Larson, 1975, Jarvis, 1985, Burwell, Jarvis, and McKnew, 1993), in concert with prescriptive models.

The focus of this dissertation is on uncertainty in response times and its role in the performance of an EMS system. The overall contribution is to further narrow the gap between the descriptive and prescriptive literature, bringing together analytic techniques and optimal location models in order to provide a tractable, yet practical foundation for operational decisions in EMS systems. Additional, more specific contributions include the extension of a method for modelling uncertainty in the availability of ambulances, the incorporation of a number of aspects of uncertainty in ambulance response time into an optimal location model, and a thorough empirical analysis of ambulance travel times. Although the research is developed specifically in relation to ambulance services, some of the models might be more generally applicable, for example, to fire or police services or to other organizations that involve spatial and temporal uncertainty in demand or that require rapid travel to service demand.

Literature Review

A surge of work in the area of emergency service operations in the late 1960's through the mid 1970's came about at the hands of two groups, the Rand Institute of New York and an interdisciplinary group of faculty and students involved in the Innovative Resource Planning in Urban Public Safety Systems (or IRP) Project at MIT. Both of these groups performed research in each of the three main areas of emergency response, fire, police, and ambulance services. Summaries of portions of the research in these areas at the Rand Institute can be found in a number of sources including The New York City-Rand Institute (1972), Chaiken (1978), and Walker, Chaiken, and Ignall, (1979). Some of the research conducted at MIT for the IRP Project is detailed in Willemain and Larson (1977), and Larson and Odoni (1981). Additional references and discussion of research within both of these groups can be found in Larson (2002). Two of the major influences for the work in this dissertation came from these projects. The first is Larson's hypercube queueing model initially developed to describe police operations and the second is the piecewise square root-linear

travel time model of Kolesar, Walker, and Hausner (1975), developed to describe the relationship between travel time and distance for fire vehicles in New York City. Both of these models, as well as other literature that motivated the work here will be outlined in this section.

Given that much of the influential research in modelling ambulance operations was done 20 or 30 years ago, it may not seem to offer a rich context for novel research. However, as discussed in Henderson and Mason (2004), there are a number of reasons that this is not the case. The primary reason is technology (and the data that accompanies it): new technology has led to changes in how the system operates, and a wealth of data makes it possible to test model assumptions more thoroughly, and to use models that require more data as input.

The intent of this section is to provide a broad summary of the literature that is most relevant to the theme of this dissertation, uncertainty in (the response times of) EMS operations, and not to provide an in depth review of all of the literature related to such operations. More detailed exposition of the literature related to the three main topics of this dissertation is provided later, in the corresponding chapters. For more general or more comprehensive reviews of literature in EMS delivery see Brotcorne, Laporte, and Semet, (2003), Swersey (1994), or ReVelle et al. (1977). The discussion here is organized along two main streams of literature; prescriptive models that for specific performance goals and operational restrictions, offer a solution in terms of a configuration of resources that satisfy the goals and restrictions, and descriptive models that for a given configuration of resources generate one or more measures of the system's performance. The prescriptive studies use mathematical optimization models that typically make strong simplifying assumptions often leaving out probabilistic components of the problem in order to maintain tractability. Although every prescriptive model contains some form of descriptive model (this is necessary in order to evaluate different configurations to come up with a solution), we use the term descriptive model here to refer to those models that try to faithfully capture enough of the

details of the operation of a system to enable reliable prediction of the impact of changes to system operation. The descriptive models, either analytic models or simulation models, can incorporate more realistic assumptions than the prescriptive models, but do not prescribe an optimal solution and can be computationally intensive. Some researchers have examined approaches that combine prescriptive and descriptive models (Batta, Dolan, and Krishnamurthy, 1989; Berman and Krass, 2001) and the work presented here aims to further narrow the gap between these two streams of research.

The literature on prescriptive models for ambulance location/allocation has developed considerably over the past thirty years. Two early models, the location set covering model, (Toregas et al., 1971) and the maximal covering location model (Church and ReVelle, 1974), provide the foundation for most of the subsequent models in the literature. The first of these models aims to minimize the number of ambulances needed to cover all of the demand. The second model aims to maximize the total demand that is covered with a given number of ambulances. Note that for both of these models a demand point is considered covered if there is an ambulance located within a pre-specified time (or distance) standard. These first models were deterministic and left out probabilistic considerations such as ambulance availability, and response time variability. For example, these models assumed that if an ambulance is located at a station, then that ambulance is always available to respond to calls in the vicinity, and did not take into account the fact that sometimes the closest ambulance would already be on a call and that another (further) ambulance might need to respond to the incoming call.

A number of models were developed to address this shortcoming. Some of these models (Daskin and Stern, 1981; Hogan and ReVelle, 1986; Batta and Mannur, 1990) are also deterministic, and provide extra or back-up coverage to deal with ambulance unavailability. Other models, such as the maximum expected covering location model (Daskin, 1983; Saydam and McKnew, 1985), are probabilistic in

that they explicitly consider the probability of an ambulance being unavailable (i.e., on a call), henceforth referred to as the busy fraction, in their formulation. These models, however, assume that the busy fraction is constant, uniform across stations, independent of the availability of other ambulances, and exogenous to the model. More recent models attempt to relax some of these assumptions (and do so by combining a more accurate descriptive model with the prescriptive model). The maximum availability location problem (ReVelle and Hogan, 1988, 1989) aims to maximize the population that has a server within a specified travel time standard with a given probability. This model builds on Daskin's maximum expected covering location model and incorporates randomness in server availability using local estimates of the server busy fraction. In order to expand the dichotomous definition of coverage the authors introduce the concept of "reliability of coverage". Use of this concept leads to relatively tractable models, but the models have the drawback that reliability of coverage may be an inadequate proxy for the operational goals that EMS systems typically use, such as maximizing expected coverage. Another drawback of this model is that the busy fractions are assumed to be the same within a region, but it is not obvious how to determine appropriate regions. Other approaches have been taken which not only explicitly consider the ambulance availability and include site-specific busy fractions, but that also have the busy fractions endogenous to the model so that they are dependent on the number of servers in use (Batta, Dolan, and Krishnamurthy, 1989; Goldberg et al., 1990b). The first of these approaches uses the hypercube queueing model to estimate the busy fractions, thus relaxing the assumption of independence, while the second uses a similar approach but does not relax the assumption of independence. The model by Goldberg et al. also incorporates uncertainty in travel times and seeks to maximize the expected number of calls reached within a given time standard, an objective that is more common in real EMS systems, based on our experience.

The literature on descriptive models for emergency response operations can be separated into two main categories, analytical models and simulation models. These are each discussed in turn in the next two paragraphs.

There have been numerous analytic models designed to estimate some aspect of the performance of emergency service systems. The most comprehensive and influential model in this area is the hypercube queueing model (Larson, 1974), which models server cooperation and dependence between servers in spatial queueing systems. This model allows, under certain assumptions, the exact calculation of server utilization (i.e., server busy fractions) and dispatch probabilities, which are the basis for many of the performance measures important for an emergency response system, from coverage to average response time. The dispatch probabilities will be defined and discussed in greater detail in the next chapter. For now, it is important to impart that they provide information about the probability of a server responding to an incoming call and that they can capture the uncertainty in server availability as well as the dependence between servers. Larson (1975), Jarvis (1985), and Burwell, Jarvis, and McKnew (1993), calculate server specific busy fractions and dispatch probabilities using approximations to the hypercube model that assume that servers are sampled randomly without replacement. Although approximate, these analyses relax some of the assumptions in the original hypercube model. Other important developments in the area of analytic models include: the piecewise square root-linear travel time vs. distance relation (Kolesar, Walker, and Hausner, 1975) mentioned above, and additional methods for estimating vehicle utilization and dispatch probabilities (Birge and Pollock, 1989, and Goldberg and Szidarovszky, 1991a-d).

One of the very first simulation models describing an ambulance system was for the Brooklyn area of New York (Savas, 1969). Since that time there have been abundant reports of simulation used in analyzing various operational decisions for EMS systems (e.g., Swoveland et al., 1973; Goldberg et al., 1990a, Henderson and Mason, 2000 and 2004; Erkut et al., 2001; and Ingolfsson, Erkut, and Budge,

2003). Such models can be useful for studying specific aspects of emergency service systems that are not amenable to analysis with analytic models or optimal location models. Additionally, such models can be used to give a more detailed analysis of solutions suggested by these other models.

The next three chapters contain the main contributions of this dissertation. The chapters are written as separate pieces of work and although they share a common theme (modelling uncertainty in emergency service response time) and have many connections between them, they can be read independently of each other. These chapters are outlined in turn in the following paragraphs.

Chapter 2 concentrates on the dispatch probabilities of each vehicle in an emergency response system, which are important quantities required for calculating various performance measures. Estimating these dispatch probabilities is complicated by four important characteristics of emergency service systems. First, such systems involve spatially distributed queues in which different regions can have different levels of demand and consequently vehicle workload can vary between stations. Second, mainly in response to this first characteristic, some stations may have multiple vehicles. Third, an important component of the service time is the time that it takes the ambulance to travel to the location of the emergency, which introduces dependence of the service time on both the location of the server and that of the demand. Finally, since demand in this type of system is random (spatially and temporally) and urgent by nature, it is essential to have cooperation among servers. Previous approximation methods for calculating dispatch probabilities are extended to account simultaneously for the possibilities of workload variation by station, multiple vehicles per station, call and station dependent service times, and server cooperation among stations.

The ambulance location/allocation model in Chapter 3 builds on many of the models described above. It is similar to a set covering formulation in that the objective is to find the minimum number of ambulances (and to locate those

ambulances) in order to provide a specified coverage. And similar to Daskin's maximum expected covering model, rather than consider coverage as an all or none phenomenon, ambulance availability is incorporated into the measure of coverage. The specific problem addressed by the model is to determine the number of ambulances needed and where these ambulances should be located to provide a specified coverage (and given a particular response time standard). In addition to integrating uncertainty in the availability of the ambulances as well as uncertainty in the travel times of the ambulances, this model explicitly incorporates the delays prior to travel that affect the response time of the ambulance.

Chapter 4 provides an in depth examination of the travel to scene component of service. A major focus is the examination of the distribution (and in particular the variability) of actual travel times for ambulances. Additionally, consideration is given to how to incorporate that distribution into methods for estimating the travel time as a function of distance and possibly other factors.

The final chapter of the dissertation provides general conclusions and possible directions for further research.

References

R. Batta, J. Dolan, and N. Krishnamurthy (1989). The Maximal Expected Covering Location Problem: Revisited. *Transportation Science* **23** 277–287.

R. Batta, and N. Mannur (1990). Covering Location Models for Emergency Situations that Require Multiple Response Units. *Management Science* **36** 16-23.

O. Berman and D. Krass (2001). Facility Location Problems with Stochastic Demands and Congestion. In *Location Analysis: Applications and Theory*, eds. Z. Drezner and H.W. Hamacher. Springer Verlag.

L. Brotcorne, G. Laporte, and F. Semet (2003). Ambulance Location and Relocation Models. *European Journal of Operational Research* **147** 451-463.

J. Birge and S. Pollock (1989). Using Parallel Iteration for Approximate Analysis of a Multiple Server Queueing System. *Operations Research* **37** 769-779.

T. Burwell, J. Jarvis, and M. McKnew (1993). Modeling Co-located Servers and Dispatch Ties in the Hypercube Model. *Computers and Operations Research* **20** 113-119.

J. M. Chaiken (1978). Transfer of Emergency Service Deployment Models to Operating Agencies. *Management Science* **24** 719-731.

R. Church and C. ReVelle (1974). The Maximal Covering Location Problem. *Papers of the Regional Science Association* **32** 101-120.

S. Cretin, and T.R. Willemain (1979). A Model of Prehospital Death from Ventricular Fibrillation following Myocardial Infarction. *Health Services Research* **14** 221-234.

M.S. Daskin (1983). A Maximum Expected Covering Location Model: Formulation, Properties, and Heuristic Solution. *Transportation Science* **17** 48–70.

M. S. Daskin, and E.H. Stern (1981). A Hierarchical Set Covering Model for Emergency Medical Service Vehicle Deployment. *Transportation Science* **15** 137-152.

M. Eisenberg, L. Bergner, A. Hallstrom (1979). Paramedic Programs and Out-of-Hospital Cardiac Arrest: I. Factors Associated with Successful Resuscitation. *American Journal of Public Health* **69** 30-38.

E. Erkut, R. Fenske, S. Kabanuk, Q. Gardiner, and J. Davis (2001). Improving the Emergency Service Delivery in St. Albert. *INFOR* **39** 416-433.

J. Goldberg, R. Dietrich, J. M. Chen, M. G. Mitwasi, T. Valenzuela, and E. Criss (1990a). A Simulation Model for Evaluating a Set of Emergency Vehicle Base Locations: Development, Validation, and Usage. *Socio-Economic Planning Sciences*, **24** 125-141.

J. Goldberg, R. Dietrich, J. M. Chen, M. G. Mitwasi, T. Valenzuela, and E. Criss (1990b). Validating and Applying a Model for Locating Emergency Medical Vehicles in Tucson, AZ. *European Journal of Operational Research* **49** 308-324.

J. Goldberg and F. Szidarovszky (1991a). A General Model and Convergence Results for Determining Vehicle Utilization in Emergency Systems. *Communications in Statistics – Stochastic Models* **7** 137-160.

J. Goldberg and F. Szidarovszky (1991b). Methods for Solving Nonlinear Equations Used in Evaluating Emergency Vehicle Busy Probabilities. *Operations Research* **39** 903-916.

J. Goldberg and F. Szidarovszky (1991c). A Model for Determining Emergency Vehicle Utilization Under an Infinite Queue and Location Dependent Service Times. Working Paper. Department of Systems and Industrial Engineering. University of Arizona.

J. Goldberg and F. Szidarovszky (1991d). Extended Models for Determining Emergency Vehicle Busy Probabilities. Working Paper. Department of Systems and Industrial Engineering. University of Arizona.

S. G. Henderson, and A. J. Mason (2000). Development of a Simulation and Data Visualisation Tool to Assist in Strategic Operations Management in Emergency Services. School of Engineering Technical Report 595, University of Auckland, New Zealand.

S. G. Henderson, and A. J. Mason (2004). Ambulance Service Planning: Simulation and Data Visualisation. To appear in *Handbook of OR/MS Applications in Healthcare*, eds. F. Sainfort, M. Brandeau, and W. Pierskalla, Kluwer.

K. Hogan, and C. ReVelle (1986). Concepts and Applications of Backup Coverage. *Management Science* **32** 1434-1444.

A. Ingolfsson, E. Erkut, and S. Budge (2003). Simulating a Single Start Station for Edmonton EMS. *Journal of the Operational Research Society* **54** 736-746.

J. Jarvis (1985). Approximating the Equilibrium Behavior of Multi-Server Loss Systems. *Management Science* **31** 235–239.

P. Kolesar, W. Walker, and J. Hausner (1975). Determining the Relation between Fire Engine Travel Times and Travel Distances in New York City. *Operations Research* **23** 614-627.

KPMG and Fitch & Associates (2000a). *Final Report: Review of City of Edmonton Emergency Medical Services – Part 1*. Prepared for Emergency Response Department, City of Edmonton, February 11, 2000.

KPMG and Fitch & Associates (2000b). *Final Report: Review of City of Edmonton Emergency Medical Services – Part 2*. Prepared for Emergency Response Department, City of Edmonton, February 11, 2000.

R.C. Larson (1974). A Hypercube Queueing Model for Facility Location and Redistricting in Urban Emergency Services. *Computers and Operations Research* **1** 67-95.

R.C. Larson (1975). Approximating the Performance of Urban Emergency Service Systems. *Operations Research* **23** 845-868.

R.C. Larson (2002). Public Sector Operations Research: A Personal Journey. *Operations Research* **50** 135-145.

R.C. Larson , and A. R. Odoni (1981). *Urban Operations Research* Prentice Hall, Englewood Cliffs, NJ.

The New York City-Rand Institute (1972). Research in 1970-1971. Operations Research **20** 474-515.

C. ReVelle and D. Bigman, D. Schilling, J. Cohon, and R. Church (1977). Facility Location: A Review of Context-free and EMS Models. *Health Services Research* **12** 128-146.

C. ReVelle and K. Hogan (1988). A Reliability-Constrained Siting Model with Local Estimates of Busy Fractions. *Environment and Planning B: Planning and Design* **15** 143–152.

C. ReVelle and K. Hogan (1989). The Maximum Availability Location Problem. *Transportation Science* **23** 192–200.

C. Saydam and M. McKnew (1985). A Separable Programming Approach to Expected Coverage: An Application to Ambulance Location. *Decision Sciences* **16** 381-398.

E. S. Savas (1969). Simulation and Cost-Effectiveness Analysis of New York's Emergency Ambulance Service. *Management Science* **15** B608-B627.

D. W. Spaite, T. D. Valenzuela, H. W. Meislin, E. A. Criss, and P. Hinsberg (1993). Prospective Validation of a New Model for Evaluating Emergency Medical Services Systems by In-Field Observation of Specific Time Intervals in Prehospital Care. *Annals of Emergency Medicine* **22** 638-645.

I. G. Stiell, G. A. Wells, V. J. DeMaio, D. W. Spaite, B. J. Field, D. P. Munkley, M. B. Lyver, L. G. Luinstra, R. Ward, (1999). Modifiable Factors Associated with Improved Cardiac Arrest Survival in a Multicenter Basic Life Support/Defibrillation System. *Annals of Emergency Medicine* **33** 44-50.

A. J. Swersey (1994). The Deployment of Police, Fire, and Emergency Medical Units. In *Operations Research and the Public Sector*, eds. S. M. Pollock, M. H. Rothkopf, and A. Barnett. North Holland, Amsterdam.

C. Swoveland, D. Uyeno, I. Vertinsky, and R. Vickson (1973). Ambulance Location: A Probabilistic Enumeration Approach. *Management Science* **20** 686-698.

C. Toregas, R. Swain, C. ReVelle, and L. Bergman (1971). The Location of Emergency Service Facilities. *Operations Research* **19** 1363-1373.

W. E. Walker, J. M. Chaiken, and E. J. Ignall, editors (1979). *Fire Department Deployment Analysis: a Public Policy Analysis Case Study / The Rand Fire Project.*

T. R. Willemain and R. C. Larson, eds. (1977). *Emergency Medical Systems Analysis*. Lexington Books, Lexington, MA.

Chapter 2: Approximating Ambulance Dispatch Probabilities

Motivation and Introduction

When considering potential operational changes for an emergency service system, often the main consideration is what impact the change will have on the response times or the coverage provided by the system. As a result, it can be critical to provide an accurate estimate of the coverage for a given system design. Although discrete event simulation is a tool that can often be useful in this regard, there are situations that call for a quick analytic solution, for example when coverage estimation is embedded in an optimization routine. An essential input for calculating coverage and other performance measures of emergency service systems is the probability that an incoming call at a particular location is served by a particular server. Uncertainty in server availability will affect these "dispatch probabilities" and the goal of this chapter is to develop a tractable procedure for modelling this uncertainty and approximating these probabilities. When these probabilities are known, many systemwide performance measures can be easily calculated, by conditioning on the location of the call and the location of the server and using the law of total probability.

To put the approximation procedure in context, consider a municipality served by a fleet of s ambulances, and with calls for service arriving according to a Poisson process at rate λ. Assume that the s vehicles are distributed among J stations, with s_j vehicles at station j. Let random variable X_j be the number of ambulances from station j that are busy (not available to take calls). Knowledge of the stationary joint distribution for $\{X_j\}$ allows the calculation of many important performance measures. This joint distribution is partly characterized by

the average busy fraction $\rho_j = E[X_j]/s_j$ for each station. It also contains information about the dependence between these random variables. In particular, used together with the dispatch policy for a given call location, this joint distribution will define the dispatch probabilities for each of the ambulances for that call location. Various procedures for estimating the average busy fractions, or the dispatch probabilities, that are related to the one here can be compared in terms of how they approximate this joint distribution. Such procedures generally make simplifying assumptions regarding one or more of the following four aspects of how the system operates:

1. **Number of vehicles per station:** It is common to assume only one vehicle per station, i.e., $s_j = 1$ for all j. This is a restrictive assumption because fixed costs of building a station or limitations on the number of available station sites may make it economical or necessary to have multiple vehicles at the same station.

2. **Average workload:** Some models assume that all vehicles have the same average utilization irrespective of the location of the vehicle's home station. This can be unrealistic because spatial variation in demand and transport network characteristics will tend to create imbalances in workload.

3. **Average service time:** Some models assume that average service time (the time a vehicle is unavailable to respond to new calls while responding to a specific call) is independent of the location of the vehicle's home station, independent of the location of the call, or both. The service time depends on the location of the call and that of the responding station due to the travel time between the two, but components other than the travel time might depend on the location of the call or responding station as well. For example certain call locations will tend to have lower transport times (if they are closer to hospitals than others) and some call locations will tend to have higher on scene times, (such as highrise apartments or office buildings).

4. **Server cooperation:** At one extreme, one could assume that each station operates as an independent subsystem, i.e., as a queueing system with s_j parallel and identical servers and some arrival rate λ_j. At the other extreme, one could assume that any call is equally likely to be responded to by any available vehicle, as in a queueing system with s servers and arrival rate λ. These two extremes simplify modeling, but reality is somewhere in between.

All of these assumptions are violated to a significant degree in real systems. For example, in the City of Edmonton Emergency Medical Services (Edmonton EMS) system, some stations can have as many as three or even four vehicles assigned to them. A detailed simulation model of this system (Ingolfsson, Erkut, and Budge, 2003) suggests that the frequency of "interdistrict dispatches" where the vehicle that responds to a call does so from a station other than the station closest to the call is about 20%, so servers from different stations do cooperate to a considerable degree. According to the model, the busy fraction varies from a low of just under 17% to a high of almost 41% across stations. The average service time varies across stations from a low of just over 43 minutes to a high of over 48 minutes.

The main contribution of the model in this chapter is the extension to site-specific (as opposed to server-specific) busy fractions and dispatch probabilities, and to allow multiple vehicles at a station. These extensions are important for a number of reasons. First, as indicated above, multiple vehicles at a station are common in real systems due to system features such as variations in the density of demand for different areas of a city, high fixed costs of opening stations, economies of scale gained by co-located servers, and the limited numbers of suitable locations for stations. Next, it is important to consider variation in the vehicle availability (rather than use a system-wide busy fraction) in order to deal not only with the possibility of multiple vehicles per station but also with spatial variation in demand and service characteristics. Additionally, it is natural to consider site-specific (rather than server-specific) availability since these characteristics vary

by location rather than by the server itself and doing so allows easy incorporation of the dispatch probabilities into optimal location models that are organized by station. Finally, in the systems that we have encountered, procedures and policies make site-specific dispatch probabilities more relevant than server-specific dispatch probabilities. For example, dispatch policies are in terms of stations rather than servers, and often servers are dynamically re-located in order to better cover demand during busy periods.

The model here addresses two limitations of existing models that do not account for the possibilities of workload variation by station, multiple vehicles per station, call and server dependent service times, and server cooperation between stations. First, existing models may be inaccurate in calculating the coverage achieved with a given number of ambulances and, conversely, may incorrectly prescribe the number of ambulances needed to meet a specified coverage objective. Second, for a given number of ambulances, existing models may prescribe a suboptimal distribution of ambulances to stations.

The remainder of the chapter is structured as follows. A brief review of the relevant literature is given next, followed by details of the approximation procedure and results of computational experiments. In a concluding section, directions for further research are discussed.

Literature Review

Two main streams of literature are relevant to the problem of considering server unavailability in emergency response systems. The first is that on the development of analytical models that allow for the calculation of measures related to server availability. The second is that related to location models for emergency service systems that incorporate such measures. Table 2-1 summarizes the methods that will be discussed in this section in terms of the four assumptions outlined in the previous section. An attempt was made to list the

models in order of increasing realism, although given the variety of assumptions, in some cases the order chosen was quite subjective.

Reference	Number of vehicles per station	Average workload	Average service time	Server cooperation
Daskin (MEXCLP, 1983)	Multiple	Constant	Constant	None
ReVelle and Hogan (1988)	Single	Allowed to vary	Constant	None
Birge and Pollock (1989)	Single	Allowed to vary	Dependent on server location and call location	None
Goldberg and Szidarovszky (1991b, c)	Single	Allowed to vary	Dependent on server location and call location	None
Goldberg and Szidarovszky (1991d)	Multiple	Allowed to vary	Dependent on server location and call location	None
Larson (Approximate Hypercube, 1975)	Single	Allowed to vary	Constant	Yes – modelled approximately
Larson (Exact Hypercube, 1974)	Single*	Allowed to vary	Dependent on server	Yes – modelled exactly
Jarvis (1985)	Single	Allowed to vary	Dependent on server location and call location	Yes – modelled approximately
Goldberg and Benitez (1990)	Single	Allowed to vary	Dependent on server location and call location	Yes – modelled approximately
Burwell, Jarvis, and McKnew (1993)	Multiple	Allowed to vary	Dependent on server location and call location	Yes – modelled approximately

Table 2-1: Summary of model assumptions of previous literature involving methods for estimating busy fractions. *Note that Larson's exact hypercube model can be extended to consider multiple vehicles per base at the cost of an enlarged state space.

A major development in the first area is the hypercube model (Larson, 1974), which models server cooperation and dependence between servers in spatial queueing systems. This model allows the exact calculation of server-specific busy fractions and dispatch probabilities and implicitly assumes that there is only one server per station. For an s server system, this model involves the solution of 2^s simultaneous equations, and as a consequence it is not practical for large systems. The assumption of a single server per station can be relaxed, at the cost of increasing the size of the state space, along with the number of equations to be solved, to $\prod_{j=1}^{s}(s_j+1)$. Larson (1975) and Jarvis (1985) calculate server-specific busy fractions and dispatch probabilities with dependence using approximations to the hypercube model that assume that servers are sampled randomly without replacement from an $M/M/s/\infty$ system (Larson, 1975) or an $M/M/s/s$ system (Larson, 1975, and Jarvis, 1985). In addition to the improvement in tractability of these approximate models, Jarvis' model allows one to consider service times that depend on the server and the customer so that variations in the portion of the service time that comprise the time for the vehicle to travel to the call location as well as in other components of the service time that could depend on the call location (for example the time spent on site or the transport time to a hospital) can be taken into account. Birge and Pollock (1989) formulate a method, similar to Larson's approximate hypercube model, in which a system of non-linear equations is solved iteratively in order to approximate a much larger exact linear equation system. Their method is not restricted to binary server states and thus is suitable for application to police systems where the server may be in various states such as on patrol, busy serving an emergency call, or busy serving a routine call. When there is more than one server located at a particular station it would usually be desirable to distribute the station's workload evenly between those servers and so these ambulances should be dispatched with equal probability to incoming calls. When two or more servers are equally preferred in the dispatch order for a particular demand location, it is referred to as a preference tie. Burwell, Jarvis, and McKnew, (1993) extend the hypercube

approximations by providing ways to account for preference ties and co-located servers. They suggest a "modified internal stacking method", that computes server-specific utilization and dispatch probabilities in the presence of arbitrary preference ties, making use of the correction factors developed by Larson (1975). Although there may be reasons for preference ties other than multiple vehicles at a station (such as two stations of equal distance from the demand location), these reasons do not seem all that common, and by focusing on multiple vehicles at a station rather than a general case of preference ties, a simpler procedure is obtained here. The main difference between the procedure described in this chapter and the modified internal stacking method proposed by Burwell, Jarvis, and McKnew, is that the dispatch probabilities (and vehicle utilizations) are calculated for each station, rather than for each server. While the modified internal stacking method uses the original correction factors developed by Larson, assuming sampling of vehicles, the procedure here is based on a new set of correction factors, for sampling of stations.

Goldberg et al. (1990) describe a method for calculating server-specific busy fractions in order to calculate expected coverage in the objective function of their optimization problem. A number of related papers present extensions to this model (including allowing co-located servers) (Goldberg and Paz, 1991, Goldberg and Szidarovszky 1991d), and provide a focus on estimation of the server busy fractions (Goldberg and Szidarovszky, 1991a-d). Many of these works include an assumption of independence between servers and in one (Goldberg and Szidarovszky, 1991d), the authors suggest that a way to improve the accuracy of the estimated busy fractions would be to include correction factors similar to those of Jarvis, but state that these had not been developed for the extensions in that paper (multiple vehicles per station and multiple vehicles responding to a call). One paper (Goldberg and Benitez, 1990) presents a method (the Decomposition method) for approximately calculating server busy fractions that does not assume independence and compares the results to the results obtained

using Jarvis' approximate method and an approximation that assumes independence between servers. The results indicated that both the Decomposition method and Jarvis' method performed better than the independence assumption method and that as the system utilization increased, the differences became more pronounced. They also found that the Decomposition method and Jarvis' method performed equally well for low utilizations, but that Jarvis' method performed better for higher utilizations. Some lessons from the papers of Goldberg and Szidarovszky (1991a-d) are relevant to the work presented in this chapter. The first is that for estimating the server utilization, a Seidel iterative process is found to converge at a faster rate than a Normal iterative process over a broad range of cases. The next is that, it is valuable to formulate the problem in such a way that the server busy fractions at each step of the iterative process will always stay in the range [0, 1]. Finally, they suggest a way to deal with incorporation of correction factors to correct for the assumption of independence, without affecting the convergence results. In particular, they were able to provide some theoretical guarantees for convergence, (i.e., a set of sufficient conditions that guarantee convergence) under the independence assumption, but could only extend these to the approximate hypercube procedure by assuming a single server at each station and that the correction factors were pre-specified constants, independent of the system utilization.

The second stream of literature, location models for emergency service systems that incorporate methods of modelling server unavailability, is relevant in particular in terms of motivating the work here. Taken together, these papers highlight the importance of modelling server unavailability and specifically, the need for models that take into account the aspects of emergency service systems considered in this chapter (demand variation by station, multiple servers per station, customer/server dependant service times, and server cooperation). An early major development in accounting for ambulance unavailability in location models was Daskin's maximum expected covering location model (MEXCLP)

(1983), which incorporates a system-wide busy fraction into the maximal covering location problem (MCLP) of Church and ReVelle (1974), in order to account for the possibility that an ambulance may not be available to respond to a call because it is busy. Batta, Dolan, and Krishnamurthy (1989) examine three assumptions of Daskin's MEXCLP model: that the busy fraction is independent of other ambulances, that it is the same for all ambulances, and that it does not depend on the location of the particular ambulance. They state that the independence assumption is not valid in systems with server cooperation, or when the servers are located in dissimilar districts (in terms of the relative amount of demand or the distribution of demand) and conclude that the expected coverage predicted by Daskin's model overestimates that calculated when considering server cooperation and allowing busy fractions to vary by server. Additional, more recent, discussions and investigations regarding solutions of the MEXCLP model are given by Aytug and Saydam (2002), and Chiyoshi, Galvao, and Morabito (2002). ReVelle and Hogan (1988, 1989) incorporate local estimates of ambulance unavailability (region-specific busy fractions). They develop a procedure to estimate these busy fractions and solve a coverage type optimization model iteratively, but find that this combined procedure does not converge and so instead reformulate the model in terms of "reliability" in order to obtain a stable solution to the problem. Unfortunately maximizing reliability is not equivalent to maximizing expected coverage in an EMS system and furthermore the determination of regions within which busy fractions should be the same is not natural or obvious. Finally, as indicated earlier, Goldberg et al. (1990) include server-specific busy fractions in calculating expected coverage in the objective function of their optimization problem.

The Approximation Procedure

Calls for service are assumed to arrive according to independent Poisson processes from a set of M demand nodes, with arrival rate λ_m from node m, and a

total arrival rate of $\lambda = \sum_{m=1}^{M} \lambda_m$. Vehicles are distributed among J stations with station j having s_j vehicles and the total number of vehicles is $s = \sum_{j=1}^{J} s_j$. The average service time for calls originating at node m served by an ambulance from station j is τ_{jm}. This includes the average travel time between station j and node m, the average time spent in service on the scene, and the average time spent in service away from the scene (transport to a hospital and time spent at the hospital). A fixed dispatch policy is assumed, where the preference of station j in the dispatch order for node m is given by a_{jm} (for example $a_{jm} = 3$ means that station j is the 3rd most preferred for responding to a call from node m).

The procedure presented here generalizes an approximation procedure for the hypercube queueing model developed by Larson (1975) and later modified by Jarvis (1985). As a starting point, we apply Little's law to the s_j servers at station j. The arrival rate to this station equals $\sum_{m=1}^{M} \lambda_m f_{jm}$, where f_{jm} is the probability that station j responds to a random call from node m. If τ_{jm} is the average service time for a call from node m that station j responds to, then the overall average service time for all calls that station j responds to is $\sum_{m=1}^{M} \lambda_m f_{jm} \tau_{jm} \Big/ \sum_{m=1}^{M} \lambda_m f_{jm}$. Little's law then implies that the average number of busy servers at station j, which can be expressed as $s_j \rho_j$, equals the total arrival rate to the station multiplied by the overall average service time for calls that the station responds to. After rearrangement, this results in the following equation:

$$\rho_j = \frac{1}{s_j} \sum_{m=1}^{M} \lambda_m f_{jm} \tau_{jm} \qquad (1)$$

The only unknown quantities on the right-hand-side of (1) are the dispatch probabilities f_{jm}. If these probabilities could be approximated as a function of

known quantities and the busy fractions ρ_j, then we would have the ingredients for an iterative procedure for estimating the busy fractions and dispatch probabilities.

To approximate the dispatch probabilities, Larson (1975) and Jarvis (1985) started with the "no cooperation" assumption. When $s_j = 1$ for all j (as both Larson and Jarvis assume) and station j is the k^{th} preferred for node m (i.e., $a_{jm} = k$), this assumption leads to approximating the dispatch probability, f_{jm}, with the product of the probabilities that ambulances at all more preferred stations are busy, multiplied with the probability that station j has a free ambulance, or:

$$f_{jm} \approx \rho_{(1)m}\rho_{(2)m}\cdots\rho_{(k-1)m}(1-\rho_j) \tag{2}$$

Here $\rho_{(l)m}$ is the busy fraction for the l^{th} preferred station for node m. To improve approximation (2), Larson and Jarvis multiplied the right-hand-side with a factor Q to approximately correct for the erroneous assumption of no cooperation:

$$f_{jm} \approx Q(s,\rho,k)\prod_{l=1}^{k-1}\rho_{(l)m}(1-\rho_j) \tag{3}$$

where ρ is an estimate of the overall system utilization (we discuss how to estimate ρ in the next section). The correction factor involves occupancy probabilities for an $M/M/s/s$ loss system, as explained later in this section. Denoting the steady state probability that the loss system has i customers by P_i, the correction factor in (3) can be expressed as

$$Q(s,\rho,k) = \frac{P_0}{s!\left(1-\rho\left(1-P_s\right)\right)}\cdot\frac{(s-k)!}{\left(1-P_s\right)^{k-1}}\cdot\left(\sum_{i=k-1}^{s-1}\frac{(s-i)\cdot s^i\cdot\rho^{i-k+1}}{(i-k+1)!}\right) \tag{4}$$

Combining equations (1), (3), and (4) leads to an iterative procedure for approximating the busy fractions and dispatch probabilities.

Larson (1975) gives a number of properties for this correction factor. To begin with, note that the correction factor can be interpreted as the relative amount by which $(1-\rho)$ overestimates or underestimates the conditional probability of the selected server being free, given that all previously selected servers are busy. First, $Q(s,\rho,0)=1$, indicating that $(1-\rho)$ does not overestimate or underestimate the probability for the first selected server. Next, $Q(s,\rho,1)<1$, indicating that $(1-\rho)$ overestimates the probability for the second selected server being free, given that the first selected server is busy. Finally, for $\rho<1-2/N$, $Q(s,\rho,k)$ is a unimodal (decreasing then increasing) function of k, and for $\rho>1-2/N$, $Q(s,\rho,k)$ is a monotonically decreasing function of k.

To allow for more than one ambulance at some stations, we generalize equations (3) and (4). The counterpart to equation (3) is

$$f_{jm} \approx Q(s,\{s_{(k)}\},\rho,m,k)\prod_{l=1}^{k-1}\rho_{(l)m}^{s_{(l)m}}(1-\rho_j^{s_j}), \qquad (5)$$

where $s_{(l)m}$ is the number of ambulances at the l^{th} preferred station for node m and we continue to assume that $a_{jm}=k$. Note that in this equation, the correction factor depends not only on s, ρ, and k, but also on how the s ambulances are distributed between stations and on the node m. This is because the preference of an ambulance will depend not only on the number of more preferred stations, but also on the number of ambulances at those stations. If the problem is constrained to allow only one ambulance per station, then the ambulance preference is the same as the station preference, but when multiple units are allowed, this is no longer the case.

We now derive an expression for the generalized correction factors $Q(s,\{s_{(k)}\},\rho,m,k)$. Consider a fictional $M/M/s/s$ system with arrival rate λ, average service time τ (we discuss how to estimate τ later), and let $\rho = \lambda\tau/s$. To simplify notation, for the remainder of this section we suppress the dependence of various quantities on the node m. We establish a correspondence between the fictional system and the real system as follows. When a call arrives from node m, the dispatcher in the real system first checks whether any of the $s_{(1)}$ ambulances at the most preferred station for that node are available. If none are available, the dispatcher checks whether any of the $s_{(2)}$ ambulances at the second most preferred station are free, and so on, until a station is found with at least one free ambulance. The corresponding sequence of events in the fictional system is to first select $s_{(1)}$ servers at random and check whether at least one of them is idle. If not, then select $s_{(2)}$ servers at random from the $s - s_{(1)}$ servers that have not been checked already (i.e., sampling without replacement) and continue in this manner until a station with at least one free ambulance is found.

With this correspondence in mind, we define the following events for the fictional system:

S_i : exactly i servers are busy
B_k : all servers at k^{th} preferred station are busy
F_k : the k^{th} preferred station has at least one free server

Additionally, we define $B_{1,n} \equiv B_1 \cap B_2 \cap \cdots \cap B_n$.

Using the law of total probability, we can express the probability that the first free server is found at the k^{th} preferred station as

$$\begin{aligned}\Pr\{B_{1,k-1} \cap F_k\} &= \sum_{i=1}^{s} \Pr\{B_{1,k-1} \cap F_k \mid S_i\} P_i \\ &= \sum_{i=1}^{s} \Pr\{B_{1,k-1} \mid S_i\} \Pr\{F_k \mid B_{1,k-1} \cap S_i\} P_i\end{aligned} \tag{6}$$

Letting $z_{(k-1)} = s_{(1)} + s_{(2)} + \ldots + s_{(k-1)}$ be the total number of ambulances at the $k-1$ most preferred stations, we can express the probability that all of these ambulances are busy, given that a total of i servers are busy, as

$$\Pr\{B_{1,k-1} \mid S_i\} = \begin{cases} 0 & \text{if k} = 1 \text{ or } z_{(k-1)} > i \\ \prod_{u=0}^{z_{(k-1)}-1} \dfrac{i-u}{s-u} & \text{if k} > 1 \text{ and } z_{(k-1)} \le i \end{cases} \tag{7}$$

The probability that the k^{th} preferred station has at least one free ambulance, given that all ambulances at the $k-1$ most preferred stations are busy and a total of i ambulances are busy is

$$\begin{aligned}\Pr\{F_k \mid B_{1,k-1} \cap S_i\} &= 1 - \Pr\{B_k \mid B_{1,k-1} \cap S_i\} \\ &= \begin{cases} 1 & \text{if } z_{(k)} > i \\ 1 - \prod_{u=0}^{s_{(k)}-1} \dfrac{i-(z_{(k-1)}+u)}{s-(z_{(k-1)}+u)} & \text{if } z_{(k)} \le i \end{cases}\end{aligned} \tag{8}$$

Combining (6) – (8) and substituting $P_i = (\rho s)^i P_0 / i!$ results, after considerable but straightforward algebra, in

$$\begin{aligned}\Pr\{B_{1,k-1} \cap F_k\} &= P_0 \sum_{i=z_{(k-1)}}^{s-1} \frac{(\rho s)^i}{i!} \prod_{u=0}^{z_{(k-1)}-1} \frac{i-u}{s-u} \left[1 - \prod_{u=0}^{s_{(k)}-1} \frac{i-(z_{(k-1)}+u)}{s-(z_{(k-1)}+u)} \right] \\ &= P_0 \sum_{i=z_{(k-1)}}^{s-1} \frac{(\rho s)^i}{i!} \left[\prod_{u=0}^{z_{(k-1)}-1} \frac{i-u}{s-u} - \prod_{u=0}^{z_{(k)}-1} \frac{i-u}{s-u} \right]\end{aligned} \tag{9}$$

Now it is necessary to relate $\Pr\{B_{1,k-1} \cap F_k\}$ to the dispatch probabilities f_{jm} of the real system. In the fictional system, the fraction of time each server is busy is $\rho(1-P_s)$.

Therefore, it makes sense to set

$$\Pr\{B_{1,k-1} \cap F_k\} = Q(s,\{s_{(k)}\},\rho,m,k)\left(\rho(1-P_s)\right)^{z_{(k-1)}}\left(1-\left(\rho(1-P_s)\right)^{s_{(k)}}\right) \qquad (10)$$

Solving for the correction factor and substituting (9), gives

$$Q(s,\{s_{(k)}\},\rho,m,k) = \frac{P_0 \sum_{i=z_{(k-1)}}^{s-1} \frac{(\rho s)^i}{i!}\left[\prod_{u=0}^{z_{(k-1)}-1}\frac{i-u}{s-u} - \prod_{u=0}^{z_{(k)}-1}\frac{i-u}{s-u}\right]}{\left(\rho(1-P_s)\right)^{z_{(k-1)m}}\left(1-\left(\rho(1-P_s)\right)^{s_{(k)m}}\right)} \qquad (11)$$

Figure 2-1 illustrates how $Q(s,\{s_{(k)}\},\rho,m,k)$ varies with ρ and k and for different scenarios $\{s_{(k)}\}$. In the first scenario (panel 1), each of ten stations has one server. In this case, (11) reduces to the correction factor formula (4) that Larson developed. The second scenario (panel 2) is identical to the first except that the fourth preferred station has two servers. The third scenario (panel 3) has two servers at the second preferred station, three servers at the fourth preferred station, and one server at the remaining stations. In panel 4, the correction factors for a number of different scenarios, or $\{s_{(k)}\}$, are shown as identified in the legend, all for $\rho = 0.4$. By comparing the graphs, one can see that increasing the number of servers at a particular station results in steeper functions beyond that station (towards the less preferred stations), and that the impact is much larger for the lower values of system utilization. It is also evident from the graph in panel 3 that as the number of servers increases, the linearly interpolated curves will not necessarily remain convex. Although the vertical axis is truncated (at a value of 4 for the first 3 panels, and a value of 30 in the fourth panel), the values of Q can be much higher than this, especially at low utilizations and when there are multiple servers in the most preferred stations (at the lower values of k). For example, for the scenario shown in the third panel, with $\rho = 0.1$ and $k = 10$, Q is 1133. However, at such a low system utilization, the correction factor for $k = 10$ is not

very relevant since the chance of finding all of the servers busy at the first 9 preferred stations (in this case, 12 servers) is extremely unlikely. An additional insight from panel 4 is that the correction factors are the same for the same number of total more preferred servers at a given k, even if the $\{s_{(k)}\}$ vector up to that k is not the same (e.g., scenarios $\{1,3,1,1,1,1,1,1,1,1\}$ and $\{1,1,1,1,1,3,1,1,1,1\}$).

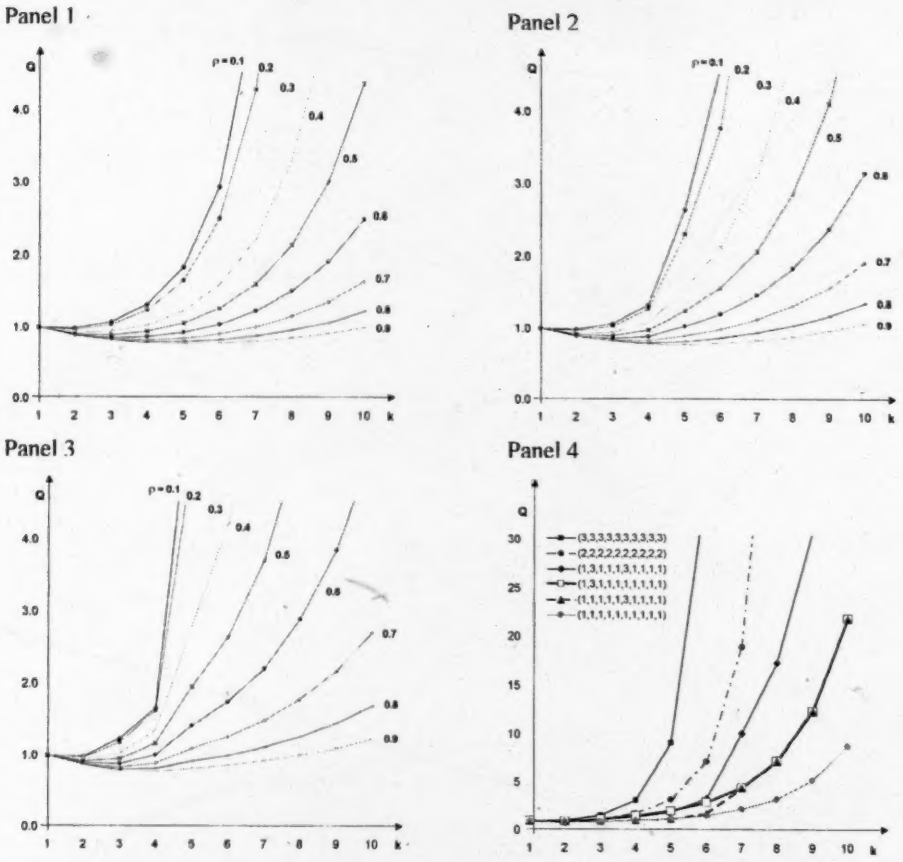


Figure 2-1: Graphs of $Q\left(s,\{s_{(k)}\},\rho,m,k\right)$. The first panel is for one server per station, the second panel has an additional server at the fourth preferred station, and the third panel has an additional server at the second preferred station and two additional servers at the fourth preferred station. Panel 4 is for $\rho = 0.4$ and gives a number of different scenarios, or $\{s_{(k)}\}$, as identified in the legend.

Equations (1), (5), and (11) provide the building blocks that we use in our algorithm to estimate station-specific busy fractions and dispatch probabilities, as we describe next.

Algorithm

The input to the algorithm is the arrival rate λ_m for every node, the number of servers s_j at each station, the preference order for each node as specified by a_{jm}, and the average service times τ_{jm}, for $m = 1, 2, \ldots, M, j = 1, 2, \ldots, J$. Stations that have no ambulances are assumed to have been removed during preprocessing, so that $s_j \geq 1$ for all j. First, calculate the following;

$$
\begin{aligned}
b_{km} &= k^{\text{th}} \text{ preferred station for node } m \\
s_{(k)m} &= \text{number of vehicles at station } b_{km} \\
z_{(k)m} &= s_{(1)m} + s_{(2)m} + \ldots + s_{(k)m} \\
\tau_{(k)m} &= \tau_{b_{km},m}
\end{aligned}
$$

Next, set the iteration counter h to 1 and initialize the busy fractions and the system-wide average service time, by assuming that all calls are responded to by the most preferred station (superscripts are used as iteration counters):

$$
\rho_j^h = \frac{1}{s_j} \sum_{m:b_{1m}=j} \lambda_m \tau_{jm}
$$

$$
\tau^h = \frac{1}{\lambda} \sum_{m=1}^{M} \lambda_m \tau_{(1)m}
$$

Each iteration consists of the following steps:

Step 1: Calculate $\rho^h = \lambda \tau^h / s$. Use ρ^h and s to calculate P_0 and P_s.

Step 2: Calculate V_j^h for all j, using (11) and the following:

$$V_j^h = \sum_{m=1}^{M} \lambda_m \tau_{jm} Q(s, \{s_{(k)}\}, \rho^h, m, k) \prod_{l=1}^{k-1} \left(\rho_{(l)m}^{h\ s_{(l)m}} \right)^h \tag{12}$$

Step 3: Calculate ρ_j^{h+1} for all j using

$$\rho_j^{h+1} = \frac{V_j^h}{s_j + \left(\rho_j^h\right)^{s_j - 1} V_j^h} \tag{13}$$

Step 4: If $\left|\rho_j^{h+1} - \rho_j^h\right| < \varepsilon$ for all j then stop. Otherwise, set $h = h+1$, $P_s = \max\left(0, 1 - \sum_{j=1}^{J} s_j \rho_j^h / (s\rho^h)\right)$, calculate f_{jm} using (5), and calculate τ^h using

$$\tau^h = \frac{1}{\lambda(1-P_s)} \sum_{m=1}^{M} \lambda_m \sum_{j=1}^{J} f_{jm} \tau_{jm} \tag{14}$$

Then return to step 1.

Computational Results

Convergence

Two datasets were used to investigate the convergence of the algorithm. The first is from Greenville County, South Carolina (Burwell, 1986). This dataset consists of 5 stations and 99 demand nodes, and 3125 scenarios (all possible combinations of allocations of 0 to 4 ambulances per station among the five stations) were run. For each scenario, the number of servers per station was allowed to vary, but it was limited to at most 4 ambulances since we felt that more than 4 ambulances per station was not reasonable. The procedure converged in all 3,125 cases and generally it took only 4 or 5 iterations to do so. The maximum number of iterations was 13 over this test set and the average was 4.18. Even more

impressive was the time to converge. For this dataset the longest time was 0.063 seconds and the average time was about 0.036 seconds.

The second dataset, from Edmonton Alberta (Ingolfsson, Erkut, and Budge, 2003), consists of 10 stations, and 180 demand nodes and over 55,000 scenarios were run. The maximum number of ambulances for each station was constrained by the actual capacity for that station (reported by Edmonton EMS). Then, all of the possible combinations over this restricted range were run. For this larger problem, the algorithm took more iterations and longer overall time to converge. For three of the 55,404 total scenarios, the procedure did not converge within the maximum allowed number of iterations (1,000). For the remaining scenarios, the number of iterations ranged from 3 to 311, and averaged just under 11. The time was about an order of magnitude more compared to the smaller Greenville County results, but as indicated in Figure 2-2, it was almost always under half a second to converge in the cases considered.

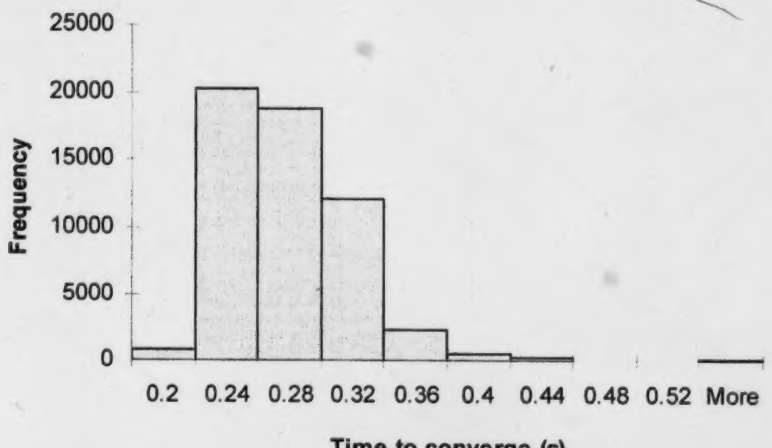


Figure 2-2: Time to converge for 55,401 scenarios of Edmonton data.

Figure 2-3 provides information about the correction factors for the subset of cases (648 in total) of the Edmonton dataset in which all stations had at least one ambulance (i.e., no empty stations). We tabulated the correction factors, $Q_{m,k}$, for these scenarios based on the value of k, the preference of the station in the response list for the demand node, using small bin sizes at the low end of the scale and larger bin sizes for higher values of $Q_{m,k}$ (and hence use a log scale for the x axes in the figure). The graphs give, for each bin, the relative portion of the correction factors for various station preferences, k. As the graphs show, in general, the stations that are very high in the preference list have smaller correction factors (stations in the top three preference positions tend to have values less than 4), and the distribution shifts to the higher values of Q, for the stations lower in the prefenace ranking.

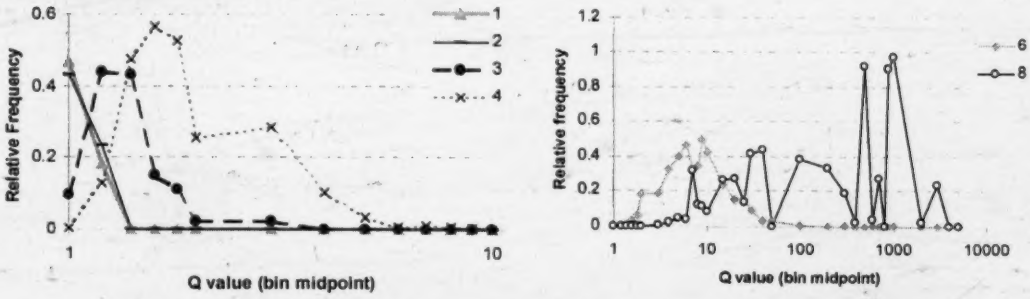


Figure 2-3: Relative frequencies of correction factors by station preference (the first through fourth preferred stations are shown on the left graph, and the sixth and eighth preferred stations are shown on the right graph) for 648 scenarios of the Edmonton dataset. Note that the x axes for both graphs use a log scale.

Once again we note that although the correction factors can be very large for the less preferred stations, these values are not really relevant under realistic system loads since the chance of needing to call on servers from those stations (for example all servers at the five most preferred stations are busy) is very small.

Accuracy

The results of the procedure were compared to the results of a discrete event simulation model for a number of scenarios using the Edmonton dataset, in order to evaluate the accuracy of the estimation procedure. For the experimental design, three different numbers of bases/stations (4, 8, and 10), and four patterns of allocations to these stations (shown in Table 2-2) were considered and each of these scenarios was run for system loads ($\rho = \lambda\tau / s$) ranging from 0.1 to 0.9. The system load was varied by changing the total arrival rate of calls to the system. Note that the values of ρ are approximate since the average service time, τ, will depend on the individual vehicle utilizations and so the average service time value was estimated (assuming an average system-wide utilization to calculate the dispatch probabilities). Additionally, the impact of temporal variation (in the demand process and vehicle allocations) on the utilization estimates was investigated.

	1	2	3	4	5	6	7	8	9	10
4S-P1	1	1	1	1						
4S-P2	1	1	2	2						
4S-P3	1	2	2	3						
4S-P4	2	2	2	2						
8S-P1	1	1	1	1	1	1	1	1		
8S-P2	1	1	1	1	2	2	2	2		
8S-P3	1	1	2	2	2	2	3	3		
8S-P4	2	2	2	2	2	2	2	2		
10S-P1	1	1	1	1	1	1	1	1	1	1
10S-P2	1	1	1	1	1	1	2	2	2	2
10S-P3	1	1	2	2	2	2	2	2	3	3
10S-P4	2	2	2	2	2	2	2	2	2	2

Table 2-2: Scenarios included in experimental design. The labels in the leftmost column are used to indicate the number of stations, followed by a pattern number.

For ease of reading, for the remainder of the section "average error" is used to mean the average (across stations) of the absolute errors and "average relative error" is used to mean the average (across stations) of the relative errors. Note also that "scenario" is used to refer to a particular number of stations and allocation pattern of ambulances to those stations.

A summary of the results for the 108 cases in the experimental design is given in Table 2-3.

ρ	4S-P1	4S-P2	4S-P3	4S-P4	8S-P1	8S-P2	8S-P3	8S-P4	10S-P1	10S-P2	10S-P3	10S-P4
0.1	0.0005	0.002	0.001	0.000	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
0.2	0.001	0.002	0.003	0.002	0.002	0.002	0.002	0.003	0.001	0.003	0.004	0.004
0.3	0.003	0.003	0.002	0.003	0.003	0.003	0.004	0.005	0.002	0.004	0.006	0.005
0.4	0.002	0.004	0.006	0.006	0.002	0.003	0.007	0.006	0.005	0.006	0.007	0.006
0.5	0.002	0.004	0.006	0.004	0.003	0.004	0.006	0.007	0.006	0.010	0.011	0.014
0.6	0.003	0.004	0.006	0.003	0.003	0.005	0.005	0.006	0.006	0.009	0.014	0.018
0.7	0.003	0.004	0.006	0.003	0.003	0.004	0.005	0.004	0.006	0.007	0.011	0.013
0.8	0.003	0.004	0.005	0.003	0.003	0.005	0.005	0.004	0.004	0.005	0.008	0.009
0.9	0.002	0.004	0.005	0.002	0.002	0.004	0.004	0.003	0.004	0.005	0.007	0.007

ρ	4S-P1	4S-P2	4S-P3	4S-P4	8S-P1	8S-P2	8S-P3	8S-P4	10S-P1	10S-P2	10S-P3	10S-P4
0.1	0.6%	1.3%	0.8%	0.3%	0.7%	0.8%	1.2%	1.6%	1.0%	1.5%	0.8%	1.5%
0.2	0.9%	0.9%	1.5%	1.1%	0.7%	1.1%	1.4%	1.7%	0.8%	1.4%	1.7%	1.9%
0.3	1.2%	1.2%	0.8%	1.1%	0.9%	0.9%	1.7%	1.7%	0.9%	1.4%	2.0%	1.9%
0.4	0.6%	1.1%	1.5%	1.7%	0.6%	0.8%	1.8%	1.6%	1.5%	1.5%	1.8%	1.9%
0.5	0.5%	1.0%	1.2%	1.0%	0.7%	0.8%	1.2%	1.5%	1.3%	2.1%	2.4%	2.9%
0.6	0.5%	0.8%	1.2%	0.6%	0.5%	0.8%	0.8%	1.0%	1.1%	1.6%	2.4%	3.0%
0.7	0.5%	0.7%	1.0%	0.5%	0.4%	0.6%	0.8%	0.6%	0.9%	1.1%	1.6%	1.9%
0.8	0.5%	0.6%	0.7%	0.4%	0.4%	0.7%	0.7%	0.5%	0.7%	0.7%	1.1%	1.2%
0.9	0.4%	0.6%	0.8%	0.4%	0.3%	0.5%	0.5%	0.4%	0.6%	0.7%	0.9%	1.0%

Table 2-3: Average errors and average relative errors for 108 scenarios (Edmonton dataset).

As evident in the table, the procedure typically gives average relative errors below 2%. The average errors and average relative errors tend to be highest for system loads in the middle of the range (and lower for very low or very high system

loads). Additionally, as the total number of stations increases, both the average error and the average relative error tend to increase slightly. An interesting finding is that the errors did not simply increase with additional ambulances without regard to the allocation of those additional ambulances between the stations. In fact, although the errors were higher when comparing pattern 4 (two ambulances per station) to pattern 1 (one ambulance per station), the errors were highest for the cases where the number of ambulances per station varied (patterns 2 and 3). This makes sense since the calculation of the correction factors for the approximation procedure uses an assumption that the workload and average service time are the same for servers at all stations, but servers further away from the center of the system will actually have higher average service times due to the travel component of the service time.

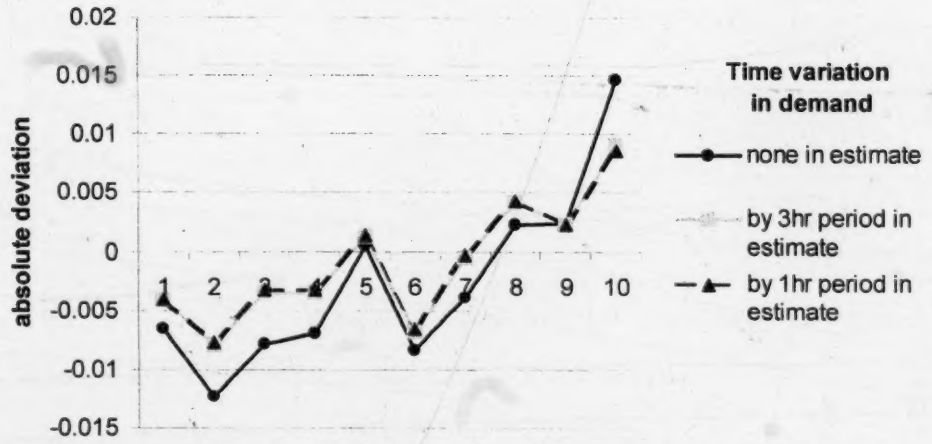
An example of the results, for a realistic scenario (with 14 ambulances allocated as in pattern 10S-P2, and system load of 0.3), is shown in the Table 2-4. The agreement between the simulated and approximated busy fractions is rather good, with most of the relative errors below 2%.

Station, j	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Average**
Servers, s_j	1	2	1	1	1	1	2	2	1	2	1.4
Estimated ρ_j	0.32	0.15	0.17	0.26	0.33	0.34	0.28	0.32	0.37	0.47	0.30
Simulation ρ_j	0.33	0.16	0.17	0.26	0.33	0.35	0.28	0.32	0.37	0.46	0.30
Absolute error	0.006	0.006	0.003	0.000	0.002	0.008	0.000	0.004	0.000	0.008	0.004
Relative error	1.9%	3.9%	1.9%	0.1%	0.8%	2.3%	0.2%	1.2%	0.1%	1.8%	1.4%

Table 2-4: Results for a particular scenario by station and averaged across stations (including estimated and simulated utilizations along with absolute and relative errors).

It is valuable to expand the analyses beyond the original experimental design detailed above, in order to examine the impact of the inclusion of certain model elements on the accuracy of the utilization estimates. The first element that we

focus on is the time varying nature of the demand. Repede and Bernardo (1994) found that modelling this element can have a large impact on the estimated performance of an EMS system. In an application to an EMS system in Louisville, Kentucky, they found that their time varying maximum expected covering location model, or TIMEXCLP, had an error 79% lower on average than the MEXCLP model (for 35 scenarios with a different number of servers and different numbers of time periods per day considered and with time variation) in comparison to results from a simulation model. Not only is it typical for the demand to vary by hour of the day or day of the week, but it is also possible to have a different number of units scheduled depending on these elements of time. First, we considered time variation in the demand only. Two of the 10 station scenarios (patterns 10S-P2 and 10S-P4) for system loads of 0.3 and 0.6 were run to examine this component. In the simulation model, a time varying arrival process was used with different rates for each hour of the day. Then time variation was included in the estimation procedure by calculating the busy fractions for each hour of the day (or for each three hour period) separately and averaging over these periods to get the final estimates. In general, the estimates for the model with time varying demand had average errors and average relative errors about twice as high compared to the same model with stationary demand. A typical example of the absolute deviations between the simulation model and the estimation procedure is shown in Figure 2-4. In all cases dividing the day into three-hour periods worked nearly as well as dividing it into one-hour periods.



Station, j	1	2	3	4	5	6	7	8	9	10
Servers, s_j	1	2	1	1	1	1	2	2	1	2
h_m, for a_{jm}=1	6%	5%	3%	4%	6%	5%	9%	10%	6%	22%

Figure 2-4: Graph of absolute error by station for a particular scenario (10S-P2, ρ=0.3, with time varying demand) for three different methods of estimating the busy fraction (without incorporating the time variation into the estimate, and incorporating it in 1-hour or 3-hour intervals). The table at the bottom shows the number of servers at each station and the proportion of demand for which that station is the closest station.

This procedure of dividing the day into smaller time periods can also be used in order to incorporate a time-varying vehicle allocation into the estimation procedure. Similar quality results were found (around 2% average relative error) when this was done for a 10 station scenario with time-varying demand and with the total number of ambulances varying from 14 to 18 depending on the time of day for both system loads tested (0.3 and 0.6).

Note that for all of the results given in this section, the procedure was modified slightly in order to make it comparable to the simulation model. In the simulation model used, an ambulance can not be considered available to respond to a call

while returning to a station (even if it is the closest ambulance to an incoming call). Thus, the time to return to a station was included as service time in our estimation procedure and in the simulation model in order to compare the results. Comparisons were also made without this modification to the input to the estimation procedure and the errors were only slightly higher in most cases.

Impact on Performance Measures

Next, some specific cases are considered in order to examine the impact of the site-specific busy fractions and correction factors on the estimated performance of the system. First, consideration is given to the busy fractions themselves. Figure 2-5 shows the solution (i.e., the estimated site-specific busy fractions) for a particular scenario of the Edmonton dataset. This scenario was based on the results of an optimization model incorporating the station-specific busy fractions using actual data from Edmonton. The horizontal axis shows the number of ambulances located at each station. Notice that the estimated busy fractions vary quite significantly by station, from a low of just over 20% to a high of nearly 60%. If accurate estimates of the busy fractions are available one can use them to calculate measures of workload imbalance (for example the difference between the highest and lowest utilizations).

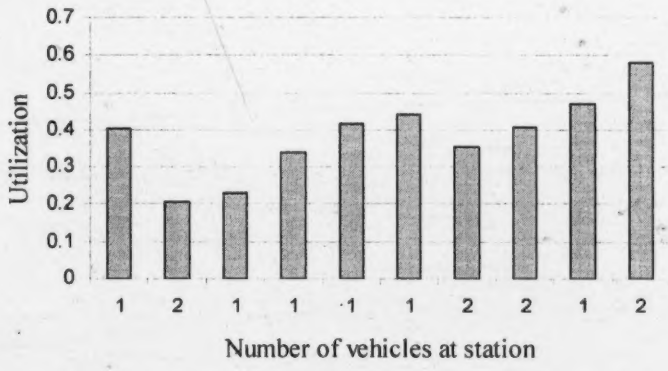


Figure 2-5: Estimated station busy fractions for a particular scenario.

Figure 2-6 shows the impact of modelling the server unavailability, and specifically includes the model with a constant system wide busy fraction as a comparison point. First of all, comparing the estimated coverage when the vehicles are assumed to always be available (i.e., using a system-wide busy fraction of zero) to that when assuming a constant system-wide busy fraction (using the average of the estimated site-specific busy fractions), the overall coverage of the system is seriously overestimated (a difference of over 6%). Next, dispatch probabilities based on a system-wide busy fraction assumption in turn overestimate the coverage of the system compared to the more realistic dispatch probabilities calculated using site-specific busy fractions and correction factors for dependence. Although the difference may seem small at 1.9%, in previous work it was found that such a difference was actually very significant in that it would require fairly major changes (for example adding two ambulances around the clock) to the system in order to attain such a difference, when the system coverage is in the vicinity of 90%. Additionally, if the coverage of smaller regions of the city are considered, the differences can be much larger. Consider the two extreme cases; the regions around the stations with the highest and lowest estimated busy fractions respectively. In the first case, the ambulances are busier than average and so the coverage for this area is over-estimated (by about 5.4%) when assuming a constant system-wide busy fraction. In the latter case, the ambulances are less busy than average and so the coverage for this area is under-estimated (by almost 8%) when assuming a constant system-wide busy fraction. This is significant because system designers may be concerned with equity in coverage between different regions, not just the system-wide coverage.

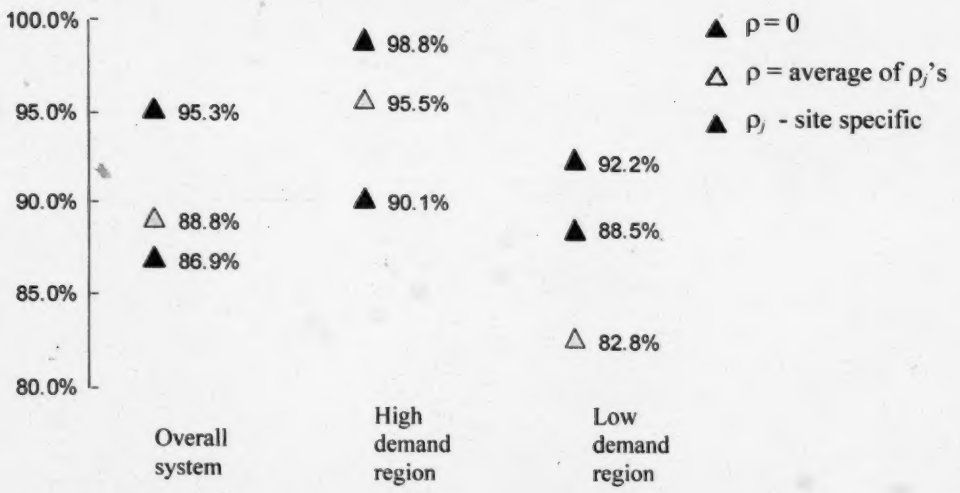


Figure 2-6: Comparison of estimated coverage for the same system using system-wide busy fraction versus site-specific busy fractions.

Other performance measures, such as average travel times, or frequency of interdistrict responses (calls responded to by an ambulance from locations other than the closest station) are easily calculated using the estimated dispatch probabilities but will not be considered in detail here.

Conclusions and Further Research

In this chapter a method for approximating station-specific vehicle utilization and dispatch probabilities for an emergency service system was detailed. Results indicate that the method is very fast, and fairly accurate for realistic scenarios, giving average relative errors of under 2% in most cases.

In the chapter that follows, this procedure is used in concert with an optimization algorithm. Of particular interest is the convergence of this combined procedure as well as the potential impact on the optimal solution. In this regard, it is beneficial that the procedure is so fast. It would be useful to extend the procedure to allow

fractional values for the number of servers at each station. This would make it possible to embed the procedure directly into an optimization procedure that solves continuous relaxations. Unless the total number of vehicles to be allocated is held constant, this extension could be quite challenging. Other extensions could include allowing for a queue of waiting calls, allowing for more than two server states (such as idle, busy on a call, or busy on patrol for police systems), and considering situations in which multiple units may respond to a call for service. Some of these extensions have been developed for previous methods (Goldberg and Szidarovszky, 1991c, d; Birge and Pollock, 1989) and it may be possible to use approaches similar to the ones used by these authors to incorporate these extensions in our model, but we have yet to investigate this.

References

H. Aytug and C. Saydam, (2002). Solving Large-Scale Maximal Expected Covering Location Problems by Genetic Algorithms: A Comparative Study. *European Journal of Operational Research* **141** 480-494.

R. Batta, J. Dolan, and N. Krishnamurthy (1989). The Maximal Expected Covering Location Problem: Revisited. *Transportation Science* **23** 277–287.

J. Birge and S. Pollock (1989). Using Parallel Iteration for Approximate Analysis of a Multiple Server Queueing System. *Operations Research* **37** 769-779.

T. Burwell (1986). A Spatially Distributed Queuing Model for Ambulance Systems. Ph.D. Dissertation, Clemson University, Clemson.

T. Burwell, J. Jarvis, and M. McKnew (1993). Modeling Co-located Servers and Dispatch Ties in the Hypercube Model. *Computers and Operations Research* **20** 113-119.

F. Chiyoshi, R. Galvao, and R. Morabito (2002). A Note on Solutions to the Maximal Expected Covering Location Problem. *Computers & Operations Research* **30** 87-96.

R. Church, and C. ReVelle (1974). The Maximal Covering Location Problem. *Papers of the Regional Science Association* **32** 101-118.

M.S. Daskin (1983). A Maximum Expected Covering Location Model: Formulation, Properties,
and Heuristic Solution. *Transportation Science* **17** 48–70.

J. Goldberg and R. Benitez (1990). Evaluating Bias in Models to Approximate Performance in Emergency Vehicle Systems. Working Paper. Department of Systems and Industrial Engineering. University of Arizona.

J. Goldberg, R. Dietrich, J. M. Chen, M. G. Mitwasi, T. Valenzuela, and E. Criss (1990). Validating and Applying a Model for Locating Emergency Medical Vehicles in Tucson, AZ. *European Journal of Operational Research* **49** 308–324.

J. Goldberg and L. Paz (1991). Locating Emergency Vehicle Bases when Service Time Depends on Call Location. *Transportation Science* **25** 264–280.

J. Goldberg and F. Szidarovszky (1991a). A General Model and Convergence Results for Determining Vehicle Utilization in Emergency Systems. *Communications in Statistics – Stochastic Models* **7** 137-160.

J. Goldberg and F. Szidarovszky (1991b). Methods for Solving Nonlinear Equations Used in Evaluating Emergency Vehicle Busy Probabilities. *Operations Research* **39** 903-916.

J. Goldberg and F. Szidarovszky (1991c). A Model for Determining Emergency Vehicle Utilization Under an Infinite Queue and Location Dependent Service Times. Working Paper. Department of Systems and Industrial Engineering. University of Arizona.

J. Goldberg and F. Szidarovszky (1991d). Extended Models for Determining Emergency Vehicle Busy Probabilities. Working Paper. Department of Systems and Industrial Engineering. University of Arizona.

A. Ingolfsson, E. Erkut, and S. Budge (2003). Simulating a Single Start Station for Edmonton EMS. *Journal of the Operational Research Society* **54** 736-746.

J. Jarvis (1985). Approximating the Equilibrium Behavior of Multi-Server Loss Systems. *Management Science* **31** 235–239.

R.L. Larson (1974). A Hypercube Queueing Model for Facility Location and Redistricting in Urban Emergency Services. *Computers and Operations Research* **1** 67-95.

R.L. Larson (1975). Approximating the Performance of Urban Emergency Service Systems. *Operations Research* **23** 845-868.

V. Marianov and C. ReVelle (1996). The Queueing Maximal Availability Location Problem: A Model for the Siting of Emergency Vehicles. *European Journal of Operational Research* **93** 110–120.

C. ReVelle and K. Hogan (1988). A Reliability-Constrained Siting Model with Local Estimates of Busy Fractions. *Environment and Planning B: Planning and Design* **15** 143–152.

C. ReVelle and K. Hogan (1989). The Maximum Availability Location Problem. *Transportation Science* **23** 192–200.

C. Saydam, and M.A. McKnew (1985). A Separable Programming Approach to Expected Coverage: An Application to Ambulance Location. *Decision Sciences* **16** 381-398.

Chapter 3: Optimal Ambulance Location With Random Delays and Travel Times

Introduction

The design of an emergency medical services (EMS) system in a municipality involves several interconnected strategic decisions, such as the number and locations of ambulance stations, the number and locations of the vehicles, the dispatch system followed, and the redeployment method used. In this chapter the focus is on the allocation of vehicles to a set of (existing or planned) ambulance stations with known locations. A main concern in an EMS system is the response time to calls. Perhaps the most obvious and significant component of response time is the travel time between the ambulance station and the demand location, and almost all of the existing operations research literature on ambulance location focuses on this component. However, since the response time is generally defined as the length of the time interval from when a call for ambulance service arrives until paramedics reach the scene, this time includes not only travel time but also delays prior to the trip. Such delays can include time spent on the phone obtaining the address and establishing the seriousness of the call, time spent deciding which ambulance to dispatch, time to contact the paramedic crew of that ambulance, and time for the paramedic crew to reach its ambulance and start it.

An overriding issue when designing an EMS system is the coverage provided, and a common performance target is to respond to (or cover) a fraction α of all calls in δ minutes or less (for example 90% in 9 minutes). A simple numerical example is offered next to illustrate the relevant issues. Consider a small town with a single ambulance station, a response time standard of 9 minutes, and three demand locations D1, D2, and D3, that are expected to generate 100 calls each in a given future time period. Suppose that the travel times between the station and the three

demand locations have means of 5.5, 7.5, and 9.5 minutes, and that the standard deviations are equal to 40% of the means. Next, suppose that the pre-trip delay is independent of the travel time and has a mean of 2.5 minutes and a standard deviation of 1 minute. Further, assume that the total response time (composed of the pre-travel delay and the travel time) follows a lognormal distribution, with a mean and standard deviation that are determined as described below. For simplicity, in this introductory example it is assumed that an ambulance is always available when a call arrives. Table 3-1 lists six different ways in which the delays and travel times can be modeled and highlights the differences among these approaches by providing results (the probability of responding to a call for each demand location, as well as the total number of calls covered) for this example.

			Probability of responding to a call at a demand location within 9 minutes			**Expected number of calls**
Model	**Travel time**	**Delay time**	**D1**	**D2**	**D3**	**covered**
A	Deterministic	Not modeled	1	1	0	200.0
B	Stochastic	Not modeled	0.929	0.747	0.521	219.7
C	Deterministic	Deterministic	1	0	0	100.0
D	Stochastic	Deterministic	0.734	0.429	0.214	137.8
E	Deterministic	Stochastic	0.857	0.129	0	98.5
F	Stochastic	Stochastic	0.708	0.426	0.229	136.3

Table 3-1: Six ways to model pre-trip delays and travel times, with summary of probabilities of responding to calls from the three demand locations for each model used, and the resulting expected number of calls covered.

If the pre-trip delay is ignored and average travel times are used to determine coverage (Model A), then the first two demand locations are characterized as "covered," the third one as "not covered," and 200 calls are credited to the coverage offered by the station when computing the performance measure. However, depending on whether and how each of the components is modelled,

the expected number of calls covered for each demand node and for the system as a whole varies widely.

Several observations regarding the differences between the models are in order.

- Comparing models A and B (or C and D, or E and F), note that the error induced by using constant as opposed to probabilistic travel times at a given demand location can be made arbitrarily large by manipulating the distance and the demand. (Suppose all of the demand is an average distance of 9.01 minutes away from the station. The deterministic model estimates zero coverage while the probabilistic model estimates over 50% coverage.) While negative and positive errors at individual demand locations may cancel each other to some extent when computing the total expected number of calls covered, the error in the system performance estimate can be quite significant (around 40% in this example when the pre-trip delays are included). We believe that the probabilistic model is a better representation of reality, and the use of deterministic travel times in ambulance location models introduces avoidable errors.
- As one would expect, the exclusion of the delay term results in very significant errors. For example, the coverage drops by more than 30% from Model B to D, due to the inclusion of the (constant) delay term.
- In the presence of probabilistic travel times, errors induced by using a constant versus probabilistic delay time are not as large as those induced by leaving out delay times altogether. Comparing Models D and F, we observe that the constant delay model (Model D) overestimates the probability for D1 by 0.026 and underestimates the probability for D3 by 0.015. Figure 3-1 displays the absolute error in the estimation of the probability (Model D probability minus Model F probability) as a function of mean travel time (in minutes) between the station and a demand point. While these errors seem small in magnitude, the relative errors can be quite significant. For example, for a travel distance of 11 minutes the absolute difference between the two probabilities is merely

0.019, but this amounts to an error of more than 13.5%. Additionally, the error will depend on the relative amount of variability in the delays compared to the travel times and it will also depend on the expected travel time as indicated in the graph. These errors can influence decisions adversely when every percent counts in trying to reach, say, a 90% coverage target. For instance, in a recent project in Edmonton (Ingolfsson, Erkut, and Budge, 2003), current coverage was 87% and most individual system design changes had impacts on the order of one percentage point or less. To be useful in such situations, prescriptive models must discriminate accurately between system designs with coverage differences of one percentage point or so.

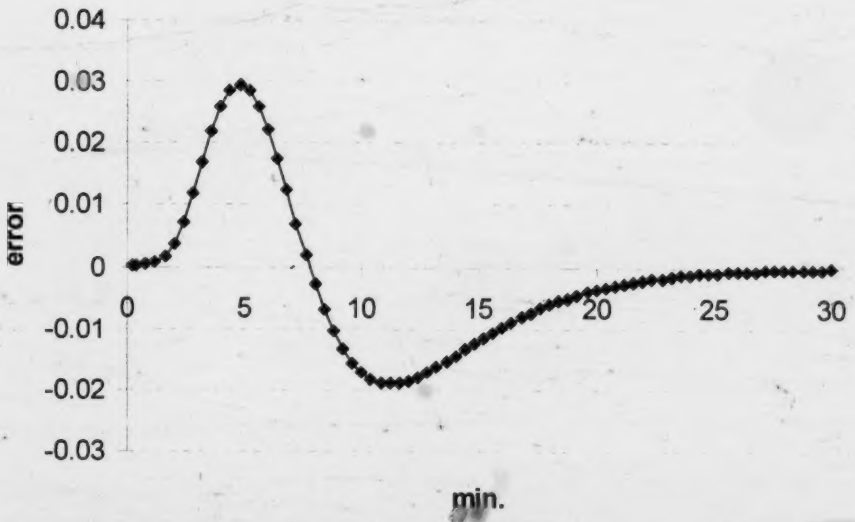


Figure 3-1: Error induced by using constant delay times as opposed to the probabilistic delay times as a function of travel time (in minutes).

In this chapter we develop Model F, which is free of the errors demonstrated in this example.

The motivation for this chapter originates from two real-world ambulance location projects completed recently – the one mentioned above and another conducted in St. Albert, a town of 50,000, near Edmonton, Alberta. Data from the latter study is used in this section. Data from approximately 5,500 EMS calls

serviced in over four years in St. Albert was analyzed. Figure 3-2 displays a histogram of the pre-trip delays. The delays ranged from 20 seconds to 20 minutes, with an average of 175 seconds and a standard deviation of 96 seconds. Limiting the analysis to calls classified as "heart and respiratory" (i.e., high priority) yielded almost the same mean and standard deviation. The average delay of almost 3 minutes is a very substantial fraction of a 9-minute response time standard, and the variation in the delay is too large to ignore (the standard deviation is almost 50% of the mean).

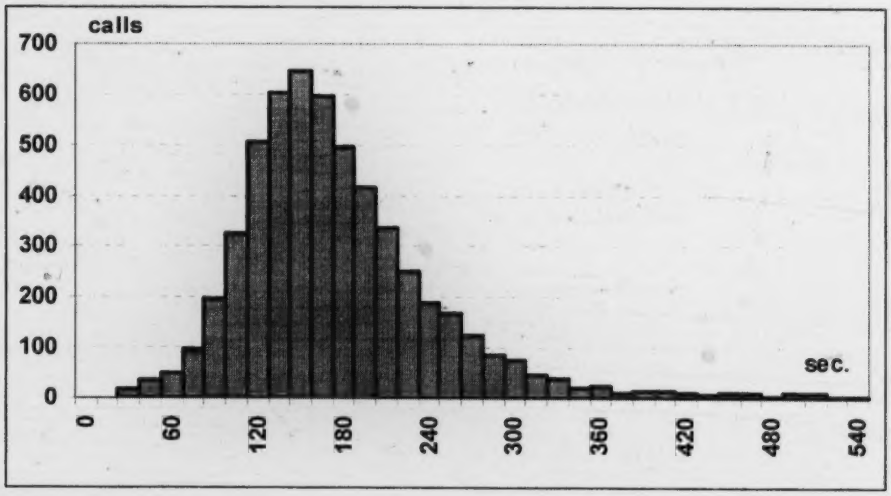


Figure 3-2: The histogram of pre-trip delays for 5,500 EMS calls serviced in St. Albert.

Green and Kolesar (1989) report delays similar to the ones of concern here. They found unexpected "dispatch delays" when validating a queueing model of police patrol in New York City. They found that about 50% of calls experienced dispatch delays averaging about 4 minutes. Henderson and Mason (2004) had a similar experience. They report that "for many of the calls, a large amount of time is spent before an ambulance is dispatched to a call" and discuss the impact that this has on the ability to meet the coverage goals as well as the potential to achieve a considerable improvement in performance with only small decreases in these pre-trip delays.

The St. Albert data set contains multiple trips to several locations, which allows analysis of distributions of travel times. Figure 3-3 shows a histogram of travel times for 352 trips from a particular station to the same multiple-resident demand point. The trip travel times range from 55 seconds to 370 seconds, with an average of 143 seconds and a standard deviation of 52 seconds. Of these 352 calls, 94 are classified as "heart and respiratory." For these high-priority calls the average travel time is 126 seconds, suggesting perhaps slightly faster travel in case of high-priority calls. However, the standard deviation is still a very substantial 57 seconds. A total of nine locations with multiple trips were analyzed and the standard deviation was always found to be considerable (on average 40% of the mean). Reporting on a project for locating emergency vehicle bases in Tucson, Arizona, Goldberg et al. (1990a) also found substantial variation in empirical travel times for given base-demand zone pairs. In Chapter 4, the focus is on the travel time component and examples that show even greater variability in travel times are provided.

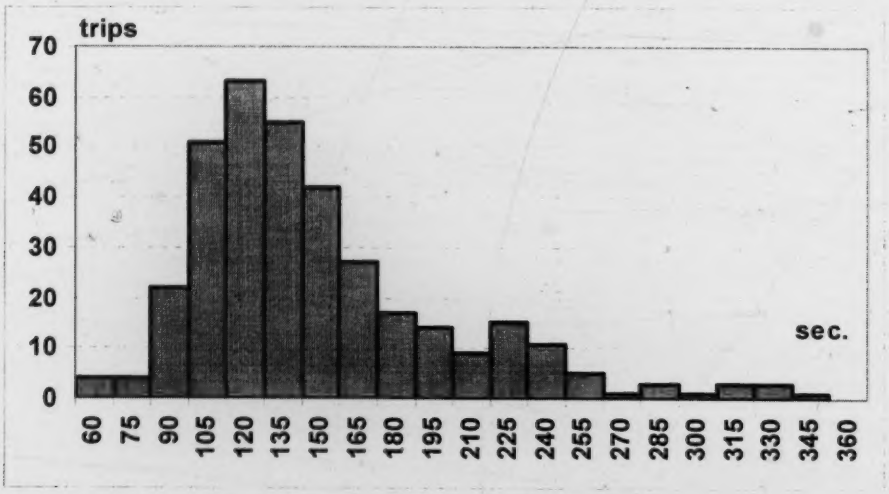


Figure 3-3: Distribution of travel times between a particular station and demand point pair for a total of 352 trips.

To summarize, in analyzing the response time data it is evident that delays can be significant and highly variable, and that travel times between a given pair of

points are highly variable. Thus, a convolution of the delay and travel time distributions is needed to obtain an accurate response time distribution (assuming travel time and delay are statistically independent – an assumption that is supported by the data used).

Explicit modelling of the uncertainty in travel times is an important feature of the model in this chapter. In addition, the model here is intended to overcome three limitations of existing models that ignore either delays or the randomness in delays. First, existing models may severely overestimate the coverage achieved with a given number of ambulances and, conversely, underestimate the number of ambulances needed to meet a specified coverage objective. Second, for a given number of ambulances, existing models may prescribe a suboptimal distribution of ambulances to stations. Third, existing models do not enable prediction of the consequences of reducing delays. This last point is important because delays can be far easier and less costly to reduce than travel times. It might be possible to reduce delays through simple process changes, such as dispatching an ambulance before the seriousness of the call has been established (thereby performing two activities in parallel rather than in series), or through the introduction of an intelligent dispatch system, whereas reducing travel times usually requires adding ambulances or stations. This model can help compare the costs and benefits of actions to reduce delays versus actions to reduce travel times. This is valuable for decision-makers who are interested in the least-costly way of reaching service standards. As far as the response time standard is concerned, 30 seconds saved are 30 seconds saved, regardless of which component of the response time these savings come from.

There is an extensive literature on optimal location of ambulances. Yet very few papers model the randomness in travel times, and we know of no papers that incorporate randomness in pre-trip delays in an optimization model. Both omissions are serious impediments to applying optimization models to ambulance

location, and the model presented here is a first step in overcoming these shortcomings.

In the remainder of the chapter, the relevant literature is discussed, and then the problem data is described, followed by the problem formulation, some useful properties of the formulation, the results of computational experiments, and further research to extend and experiment with the model.

Literature Review

There is an extensive literature on locating emergency facilities. Willemain and Larson (1977), Swersey (1994), and Marianov and ReVelle (1995) provide reviews of this area. A recent review of the literature on facility location with stochastic demands by Berman and Krass (2001) is also very relevant. In this section, selected papers are surveyed with an emphasis on those that are most relevant to this research. Past models can be characterized as prescriptive or descriptive. Every mathematical model of EMS operations provides predictions of performance, as a function of decision variables such as the number of ambulances at each station, and every such mathematical model allows one to experiment with the decision variables to search for a better configuration. All models make simplifying assumptions, for various reasons. At one extreme are models that make strong simplifying assumptions in the interest of making it possible to find optimal or near-optimal configurations for large problem instances. At the other extreme are models whose focus is on accurately predicting the performance for a particular configuration. Even though some models fall in the middle between these two extremes, many models can be usefully classified as either prescriptive (where the focus is on making optimization possible) or descriptive (where the focus is on accurate prediction of performance measures). Descriptive models are typically either analytical queueing models or simulation models.

Related to the discussion of prescriptive and descriptive models is problem size. For ambulance location models, the number of "demand nodes" and the number of stations are the primary determinants of problem size. Demand is typically aggregated into demand nodes, in part to provide a reasonable size problem. The number of demand nodes is influenced by the size of the geographic region, the population, and the method used to divide the region into demand nodes, i.e., the aggregation method. The number of stations is influenced by the size of the region, the size of the population, the level of funding, and by operating policies (for example, if it is possible to have ambulance waiting on street corners for the next call, then there would be more possible "stations"). Both the number of demand nodes and the number of stations will influence the time to evaluate a single solution, but only the number of stations (and not the number of demand nodes) will influence the size of the solution space for a prescriptive model. Moreover, the number of stations will impact the size of the problem for a prescriptive model in a combinatorial fashion. The number of demand nodes, can be expected to impact the solution time for a single solution in an approximately linear fashion. Since the number of demand nodes for a particular problem can be manipulated by using different aggregation levels (at the expense of introducing aggregation errors), and since the increases in computer performance over time can make it easier to deal with problems with increased numbers of demand nodes, the true test of a prescriptive model is the number of stations.

Most of the prescriptive models use an all-or-none notion of coverage, where a demand point is considered "covered" if the closest ambulance station is within some specified maximum distance. The objective of the set-covering location problem (SCLP), first formulated by Toregas et al. (1971), is to minimize the number of stations such that all demand points are covered. Although this is a binary problem, the linear programming relaxation (or the addition of a simple cutting plane) usually generates all-integer solutions. By changing the coverage distance, one can generate a number of solutions with varying number of facilities. While the SCLP has been used in several location studies, it has a

number of shortcomings. For example, the requirement of covering every demand point is rather stringent and usually results in the location of an unreasonably high number of facilities. To address this problem, Church and ReVelle (1974) extended the SCLP by proposing the maximal covering location problem (MCLP) where the goal is to maximize the proportion of the demand covered with a fixed number of facilities. The linear programming relaxation of this binary problem is reported to result in all-integer solutions most of the time. One can solve MCLP parametrically in the number of facilities and obtain a cost-coverage tradeoff curve. Unlike SCLP, MCLP differentiates between demand points based on relative demand and it is able to trade off system coverage and resources. Hence, it is better suited for emergency service facility location than SCLP, and there are several reported applications. However, the classification of a demand point that is within a specified distance of a station as covered makes the implicit assumption that there is always a vehicle at the station to respond to a call. While most emergency response systems are designed for low utilization levels, in many cities ambulances are busy a significant portion of the time (for example, 30%). To account for the potential unavailability of ambulances, Daskin (1983) extended MCLP by formulating the maximum expected covering location model (MEXCLP), which maximizes the expected value of population coverage for a fixed number of servers. MEXCLP uses a single, system-wide busy probability, and computes the probability of a subset of busy vehicles from a given station using the binomial distribution. While the model is an integer program with a nonlinear objective function, it can be linearized, and instances of realistic size can be solved with general-purpose integer programming solvers. Although there are many prescriptive ambulance location models in the literature, the three models discussed above can be considered the most influential ones on subsequent research, since most other models are extensions of these three. While many of these prescriptive models can be solved to optimality with reasonable effort, they suffer from simplifying assumptions. On the other hand, descriptive models provide more realism (but they do not prescribe a solution).

The main descriptive model that is relevant to this work is the hypercube model developed by Larson (1974) and subsequent approximate versions of that model (Larson, 1975 and Jarvis, 1985). The hypercube model allows busy fractions to vary between ambulances, can accommodate ambulances responding to calls outside their assigned districts, and can account explicitly for queued calls. Applications and discussions of extensions to the hypercube model are discussed in Larson (1979), and Brandeau and Larson (1986). An extension to these models, described in the previous chapter, that allows multiple servers at a station, is used here. Discrete event simulation can be used when even greater realism is needed (e.g., Henderson and Mason, 2000 and 2004, and Ingolfsson, Erkut, and Budge, 2003).

Finally, some authors have combined descriptive models with optimization heuristics. Both Batta, Dolan, and Krishnamurthy, (1989) and Saydam and Aytug, (2003) combine the approximate hypercube model with optimization heuristics, the former using a single node substitution heuristic and the latter using a genetic algorithm.

In the work described in this chapter, the prescriptive modelling paradigm is extended, by incorporating randomness in response times without sacrificing the ability to use general-purpose solvers to find optimal solutions. In the context of Table 3-1, all of the prescriptive covering models discussed above use deterministic (average) travel times. While delays are usually not explicitly mentioned in papers dealing with prescriptive coverage models, it is easy to incorporate a constant (average) delay into all coverage models by simply subtracting the delay from the specified maximum response time. For example, Eaton et al. (1985) use MCLP with a 5-minute travel time, which may have been part of an 8-minute response time with an average delay of 3 minutes. Hence, depending on the coverage standard used, SCLP, MCLP, and MEXCLP fall in category A or C in Table 3-1.

The assumption made by early covering models is that if (and only if) an ambulance is available within a specified maximum distance of a demand point, then the demand point is covered. A common performance measure for EMS systems is the coverage, or the fraction of calls responded to within a specified time standard. However, for a given ambulance location and a demand point, it is not possible to know with certainty whether the call will be responded to within the time standard – it depends on the pre-trip delay and the travel time as well as the availability of the ambulance, none of which can be predicted with certainty. The model here does not rely solely on average travel times, and hence is not limited by the resulting strict classification of demand points as covered or not covered. It allows incorporation of randomness in pre-trip delays and travel times, and computes an expected coverage for each demand point, given the ambulance locations. Hence, model realism is increased by replacing the 0-1 consequences, implied by solutions of early covering models for demand points by real numbers between 0 and 1, which are better estimates (than 0 or 1) of the fraction of calls emanating from different demand points that can be reached within the specified time standard.

In the remainder of this section, the focus is on ambulance location models that incorporate response time variability. As mentioned above, a constant pre-trip delay can be incorporated into all covering models. However, we know of no papers in the literature that incorporate random delays in a prescriptive model. We are aware of three instances where travel time variability is included in covering models. Marianov and ReVelle (1996) assume travel time from station j to node i is normally distributed with known mean and variance. Then they define a node i to be covered by station j if the average travel time plus K standard deviations is less than a specified constant. While they acknowledge the variability in travel times, they do not use the distributions directly in the model. This model is more conservative (for $K > 0$) than a covering model that uses the average travel times only. However, it is still a traditional covering model in the sense that a demand point is either covered or not. Perhaps the paper that is most

relevant to the model in this chapter is Goldberg and Paz (1991), which is inspired by a case study reported in Goldberg et al. (1990a, b). They formulate an emergency facility location model that includes the probability P_{ij} that an ambulance at station j can travel to a call from demand node i within a response time standard. This quantity is used to calculate expected coverage in the objective function of their optimization problem. Daskin (1987) models random travel times similarly, but the focus of his model is the integration of location and routing, taking into account that some calls may require two vehicles to respond. Daskin's model does not account for ambulance unavailability and is quite large, even for small networks. Goldberg and his co-workers used an approximation related to the hypercube model to estimate the busy probabilities of the vehicles, and included an upper bound on the number of stations. They use regression to estimate average travel times as a function of distance along roads of various types, and compute the P_{ij} values using this mean and the standard deviation of the residuals, assuming normal distribution of path travel times. While the way that expected coverage is modeled in this chapter is similar to that of Goldberg and Paz (1991), there are several differences between their work and the work presented here. Perhaps the most significant modelling difference is the inclusion of pre-trip delays in the current model. Also, the calculation of the dispatch probabilities for the vehicles and the computation of coverage probabilities for demand points, are treated in different ways. Here, dispatch policies are considered as given, rather than included as decision variables. For all of these reasons, the current model is more compact and tractable and problems of realistic size can be solved optimally using off-the-shelf solvers, while Goldberg and Paz (1991) propose pairwise interchange heuristics for their model.

Problem Data

The following data are assumed to be available:

- A set of J station locations, indexed by j, and a set of M demand nodes, indexed by m.
- A positive arrival rate λ_m for each demand node m. The node arrival processes are assumed to be independent Poisson processes. The system wide arrival rate is denoted with $\lambda \equiv \sum_{m=1}^{M} \lambda_m$ and the fraction of the total demand coming from demand node m is denoted by $h_m \equiv \lambda_m / \lambda$.
- The distribution function $H_{jm}(t)$ of the travel time T_{jm} from station j to node m.
- The distribution function $F(t)$ for the delay.
- Parameters δ and α which specify the coverage objective that calls should be responded to in at most δ time units with probability of at least α.
- The average on-scene time, and average time spent traveling to and remaining at a hospital, denoted $E[T_{\text{on scene}}]$, and $E[T_{\text{hospital}}]$, respectively.
- A dispatch order for each demand node m, i.e., a list of the J stations in order of preference for dispatching to a call originating from node m. The notation $k(j, m)$ is used for the preference position of station j for a call from node m, for example $k(3, 2) = 4$ indicates that station 3 is the fourth preferred station for responding to calls from node 2.
- The "busy fraction" ρ_j for ambulances at station j, i.e., the probability that an ambulance at station j is not available to respond to calls. It is assumed that $\rho_j \in (0,1)$. Additionally, correction factors, Q_{jm}, to account for the dependence between servers are assumed to be available for each station demand node pair. Details as to how these are calculated are provided in Chapter 2 of this dissertation. Together, the dispatch policy, the busy probabilities, and the correction factors are used to calculate dispatch probabilities.

The last assumption, that the busy fractions and correction factors are exogenous input to the model, is obviously a limiting one. Discussion of how to overcome this assumption is provided later.

Suppose that an ambulance from station j responds to a call from demand node m and that station j is the k^{th} station in node m's dispatch order. Let $H_{(k)m}(t)$ be the travel time distribution function for this station-demand node pair. Note that to minimize confusion, brackets are used in the station index when referring to the preference of a station as opposed to the station number. For example $H_{(k)m}(t)$ is equivalent to $H_{jm}(t)$ if station j is the k^{th} station in node m's dispatch order. With the assumed data, the probability $w_{(k)m}$ that the call will be responded to in δ time units or less, can be calculated as follows:

$$w_{(k)m} = \int_{u=0}^{\delta} H_{(k)m}(\delta - u)dF(u) \qquad (1)$$

This calculation can be done for all node-station pairs, before solving the optimization problem posed in the next section. The optimization model requires no information about the probability distributions of travel times or delays other than the probabilities $w_{(k)m}$.

The dispatch order for each node m is assumed to be such that:

$$w_{(1)m} \geq w_{(2)m} \geq \ldots \geq w_{(J)m} \qquad (2)$$

That is, the stations are arranged in descending order of the likelihood of responding to a call from node m in less than δ time units. The formulation presented in the next section is valid without this assumption, but the concavity property discussed later requires it.

Problem Formulation and Properties

Let x_j be the number of ambulances located at station j, and let $x_{(k)m}$ be the number of ambulances at the k^{th} preferred station for demand node m. The vector $(x_{(1)m}, x_{(2)m}, \ldots, x_{(J)m})$ is a permutation of $\mathbf{x} = (x_1, x_2, \ldots, x_J)$, for each m. Similarly, let $\rho_{(k)m}$ be the busy probability for the k^{th} most preferred station for demand node m. The optimization problem is:

(P1) minimize $$s(\mathbf{x}) \equiv \sum_{j=1}^{J} x_j$$

subject to $$c(\mathbf{x}) \equiv \sum_{m=1}^{M} h_m c_m(\mathbf{x}) \geq \alpha \qquad (3)$$

$$x_j \geq 0, \text{ integer, for } j = 1, 2, \ldots, J \qquad (4)$$

where

$$c_m(\mathbf{x}) = \sum_{j=1}^{J} f_{jm}(\mathbf{x}) w_{jm}, \text{ for } m = 1, 2, \ldots, M \qquad (5)$$

and

$$f_{jm}(\mathbf{x}) = Q_{jm}\left(1 - \rho_{jm}^{x_j}\right) \prod_{l=1}^{k(j,m)-1} \rho_{(l)m}^{x_{(l)m}}, \text{ for } j = 1, 2, \ldots, J,\ m = 1, 2, \ldots, M \qquad (6)$$

Problem (P1) minimizes the total number of ambulances $s(\mathbf{x})$ subject to a coverage constraint (3). Constraint (3) expresses the system-wide coverage $c(\mathbf{x})$ as a weighted combination of the coverages for each demand node, and the coverage $c_m(\mathbf{x})$ for demand node m is calculated in (5) by conditioning on which station sends an ambulance to respond to a call from node m. The calculation of the node m coverage requires the dispatch probability $f_{jm}(\mathbf{x})$, the probability that a call from node m is responded to by an ambulance from station j, its k^{th} preferred

station. This probability is calculated, as shown in (6), as the product of a correction factor, Q_{jm}, and the probabilities that all ambulances at the $k-1$ more preferred stations are busy, and at least one ambulance at station j is free. In equation (6), Q_{jm} is a correction factor to approximately account for the dependence between servers. Setting the correction factors to one is equivalent to assuming that the probability of an ambulance at a particular station being busy is statistically independent of the status of ambulances at all other stations.

Concavity Result

Proposition 1: If $w_{(1)m} \geq w_{(2)m} \geq \ldots \geq w_{(J)m}$ for $m = 1,2,...,M$, and Q_{jm} and ρ_j are invariant with **x** (recall that these are assumed to be exogenous input to the model), then the system-wide coverage is a concave function of **x**.

Proof: Recall that the system–wide coverage $c(\mathbf{x}) = \sum_{m=1}^{M} h_m c_m(\mathbf{x})$ is a convex combination of the coverage $c_m(\mathbf{x})$ for each demand node m. To prove that $c(\mathbf{x})$ is concave, it suffices to prove that the coverage $c_m(\mathbf{x})$ for a particular node m is concave, since the weights h_m are positive. Therefore, it is assumed without loss of generality that there is only one demand node and the demand node subscript m is dropped in the proof to simplify notation. Additionally, it is assumed that the stations are ordered by preference of this single demand node, so the bracket notation in the subscripts is also dropped here.

By assumption we have $\Delta w_k = w_{k+1} - w_k \leq 0$ for all k. We can express the probability $f_k(\mathbf{x})$ as:

$$f_k(\mathbf{x}) = Q_k\left(1-\rho_k^{x_k}\right)\prod_{l=1}^{k-1}\rho_l^{x_l} = Q_k\left(\prod_{l=1}^{k-1}\rho_l^{x_l} - \prod_{l=1}^{k}\rho_l^{x_l}\right) = g_{k-1}(\mathbf{x}) - g_k(\mathbf{x})$$

where $g_k(\mathbf{x}) = Q_k \prod_{l=1}^{k} \rho_l^{x_l}$ and $g_0(\mathbf{x}) = 1$. Consequently,

$$c(\mathbf{x}) = \sum_{k=1}^{J} f_k(\mathbf{x}) w_k = \sum_{k=1}^{J} g_{k-1}(\mathbf{x}) w_k - \sum_{k=1}^{J} g_k(\mathbf{x}) w_k$$
$$= \sum_{k=0}^{J} g_k(\mathbf{x}) w_{k+1} - \sum_{k=1}^{J} g_k(\mathbf{x}) w_k = w_1 + \sum_{k=1}^{J} g_k(\mathbf{x}) \Delta w_k$$

with the understanding that $w_{J+1} = 0$.

The gradient of $c(\mathbf{x})$ with respect to $\mathbf{x}$ has the following entries:

$$\frac{\partial c}{\partial x_n} = (\ln \rho_n) \sum_{k=n}^{J} g_k(\mathbf{x}) \Delta w_k$$

The Hessian matrix $\mathbf{H}$ is symmetric and has the following entries:

$$h_{ln} = h_{nl} = \frac{\partial^2 c}{\partial x_n \partial x_l} = (\ln \rho_n)(\ln \rho_l) \sum_{k=l}^{J} g_k(\mathbf{x}) \Delta w_k$$

Recalling that $Q_k > 0$, $\rho_k \in (0,1)$, and $\Delta w_k \le 0$, we see that $\partial c / \partial x_n$ is non-negative for all n, and $\partial^2 c / \partial x_n \partial x_l$ is non-positive for all n and l.

Consider the quadratic form $\mathbf{y}^T \mathbf{H} \mathbf{y}$ where $\mathbf{y}$ is an arbitrary column vector with J elements. This quadratic form can be expressed as:

$$\mathbf{y}^T \mathbf{H} \mathbf{y} = \sum_{n=1}^{J} \sum_{l=1}^{J} y_n y_l h_{nl} = \sum_{l=1}^{J} y_l^2 h_{ll} + 2 \sum_{n=1}^{J} \sum_{l=n+1}^{J} y_n y_l h_{nl}$$

Substituting the expression for h_{nl} we get:

$$\mathbf{y}^T \mathbf{H} \mathbf{y} = \sum_{l=1}^{J} y_l^2 (\ln \rho_l)^2 \sum_{k=l}^{J} g_k(\mathbf{x}) \Delta w_k + 2 \sum_{n=1}^{J} \sum_{l=n+1}^{J} y_n y_l (\ln \rho_n)(\ln \rho_l) \sum_{k=l}^{J} g_k(\mathbf{x}) \Delta w_k \quad (7)$$

By changing the order of summation, the double sum in (7) can be expressed as:

$$\sum_{l=1}^{J} y_l^2 (\ln \rho_l)^2 \sum_{k=l}^{J} g_k(\mathbf{x}) \Delta w_k = \sum_{k=1}^{J} g_k(\mathbf{x}) \Delta w_k \sum_{l=1}^{k} (\ln \rho_l)^2 y_l^2$$

Similarly, the triple sum in (7) can expressed as:

$$\sum_{n=1}^{J} \sum_{l=n+1}^{J} y_n y_l (\ln \rho_n)(\ln \rho_l) \sum_{k=l}^{J} g_k(\mathbf{x}) \Delta w_k = \sum_{n=1}^{J} \sum_{k=n+1}^{J} g_k(\mathbf{x}) \Delta w_k \sum_{l=n+1}^{k} y_n y_l (\ln \rho_n)(\ln \rho_l)$$

$$= \sum_{l=2}^{J} g_k(\mathbf{x}) \Delta w_k \sum_{n=1}^{k-1} \sum_{l=n+1}^{k} y_n y_l (\ln \rho_n)(\ln \rho_l)$$

Substitution in (7) results in:

$$\mathbf{y}^T \mathbf{H} \mathbf{y} = \sum_{k=1}^{J} g_k(\mathbf{x}) \Delta w_k \left\{ \sum_{l=1}^{k} (\ln \rho_l)^2 y_l^2 + 2 \sum_{n=1}^{k-1} \sum_{l=n+1}^{k} (\ln \rho_n)(\ln \rho_l) y_n y_l \right\}$$

$$= \sum_{k=1}^{J} g_k(\mathbf{x}) \Delta w_k \left(\sum_{l=1}^{k} (\ln \rho_l) y_l \right)^2$$

Each term in the outer summation is non-positive (because $g_k(\mathbf{x}) \geq 0$, $\Delta w_k \leq 0$, and the squared summation is non-negative) and therefore $\mathbf{y}^T \mathbf{H} \mathbf{y} \leq 0$ for all y. Consequently, **H** is negative semi-definite and $c(\mathbf{x})$ is concave.

Q.E.D.

The objective function and constraints (4) in (P1) are linear. The constraint (3) is concave, and it defines a convex set in **x**. Consequently, the continuous relaxation of (P1) is a convex programming problem, and a local optimum is also global. Note that as a result of this proposition, the coverage $c_m(\mathbf{x})$ for each demand node m has the following properties:

- An increase in the number of ambulances at any station increases the coverage for each demand node.
- When the number of ambulances at a particular station is increased, the marginal increase in coverage decreases.

Busy Fractions

The assumption that the busy fractions ρ_j are exogenous input is not realistic, as they will depend on the number and distribution of ambulances between stations. To overcome this limitation, we propose iterating between solving (P1) and estimating the busy fractions. If all ambulances are assumed to have the same busy fraction, then a relatively simple estimation procedure can be used (refer to Appendix 1 at the end of this dissertation for details). If all ambulances are not assumed to have the same busy fraction, then a more complicated estimation procedure is necessary. We have used the procedure detailed in the previous chapter, which is a generalization of the approximate hypercube model allowing for multiple vehicles at a station and estimates station-specific ambulance busy fractions.

The following iterative algorithm is proposed to overcome the assumption of the busy fractions being exogenous input.

Step 0: Set ρ_j to an initial estimate ρ_j^{in} of the busy fraction. Set $n \leftarrow 1$ and choose a smoothing parameter $\gamma \in (0,1)$.

Step 1: Solve (P1). Let the optimal objective function value be s^*. Find the solution $\mathbf{x}_n^*$ that maximizes $c(\mathbf{x})$ subject to $\sum_{j=1}^{J} x_j \le s^*$ and (4). If the convergence criterion is satisfied, stop.

Step 2: Estimate ρ_j^{out} using equation (8). Set $\rho_j^{in} \leftarrow \gamma\rho_j^{out} + (1-\gamma)\rho_j^{in}$ for all stations j and $n \leftarrow n+1$. Go back to step 1.

The convergence criterion could be expressed in terms of the sequence of solutions $\{\mathbf{x}_n^*\}$, the estimated busy fractions $\{\rho_j^{out}(\mathbf{x}_n^*)\}$, or both. This algorithm is not guaranteed to converge to a unique solution. Indeed, we have observed convergence to a cycle of two or even three similar solutions. In such cases,

planners could be presented with multiple good solutions, which could be compared in terms of the values that they give for other performance measures.

Goldberg et al. (1991) use a different approach, where they include the busy fractions as decision variables and include a constraint in the problem formulation that is similar to equation (12) in Appendix 1. An advantage of the approach in this chapter is that the continuous relaxation of (P1) is a convex optimization problem, as shown above. Goldberg et al. (1991) do not solve their formulation as a mathematical program, but use specialized heuristics.

Computational Experiments

The instances of (P1) solved in this section are based on data from Edmonton EMS and use correction factors equal to one and deterministic travel times in order to isolate the effect of randomness in delays. The dispatch orders have satisfied assumption (2). These instances have 10 stations and 180 demand nodes. These instances have been solved to optimality using Premium Solver (Frontline Systems, Inc), in at most a few minutes per instance with a standard branch-and-bound algorithm that calls a nonlinear programming algorithm to solve the continuous relaxations. We have experimented with minimizing the number of ambulances needed to provide a specified coverage (this is the formulation (P1)), as well as a formulation that maximizes the coverage subject to a given total number of ambulances.

To overcome the assumption of ρ_j being given exogenously, the values of ρ_j are iterated on using the procedure described above. Figure 3-4 shows an example of how ρ_j^{in} and ρ_j^{out} evolved over three iterations for one problem instance based on Edmonton data. In this instance, γ was set to 0.9, and ρ_j^{in} and ρ_j^{out} converge in about 3 iterations with an average after convergence of about 0.33. The total number of ambulances converged to 16. In a simulation model of the Edmonton

EMS system reported on in Ingolfsson, Erkut, and Budge (2003), the busy fraction under current operations was estimated as 0.22. In the simulation, the number of ambulances in service varied between 14 and 20, depending on the time of day, and the coverage was about 89%. The lower average utilization from the simulation model can be explained at least in part by the larger number of ambulances than in the solution illustrated in Figure 3-4.

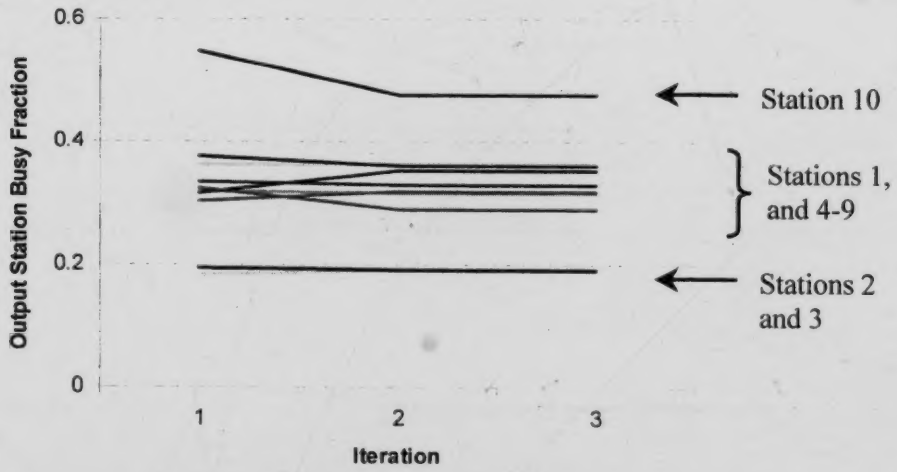


Figure 3-4: An example of iterating on the busy fractions ρ_j, where the initial input busy fraction was set to 0.3 for each station, and a smoothing constant of 0.9 was used.

The model has been used to empirically explore the impact of varying the parameters of the delay distribution. Figure 3-5 shows how the minimum total number of ambulances needed to provide the specified coverage changes when the mean and standard deviation of the delay distribution vary. Values that were 0%, 50%, 100%, 125%, and 150% of the current value for the mean (2.6 minutes) and for the standard deviation (1.3 minutes) were tried, except for combinations of parameters that made it impossible to meet the coverage goal. The combination where both the mean and the standard deviation equal their current values is referred to as the *base case*.

As Figure 3-5 shows, the total number of ambulances needed changes considerably when the parameters of the delay distribution are varied. The dramatic impact of ignoring the delay is illustrated by comparing a case when the delay is assumed to be zero to the base case. In the former case, only 11 ambulances are needed, while in the base case, 16 are needed.

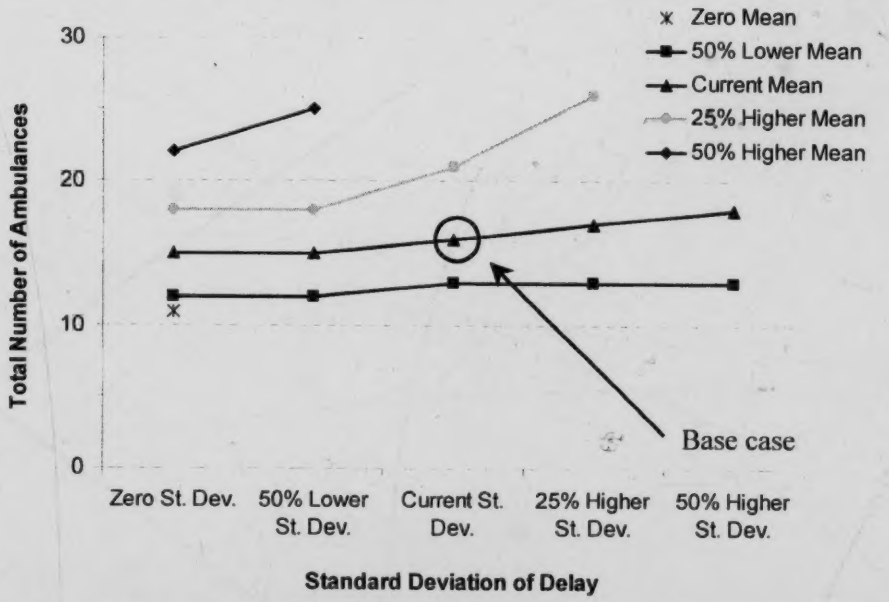


Figure 3-5: Sensitivity of the minimum total number of ambulances needed to provide the coverage goal to the mean and standard deviation of the delay distribution.

Comparison of the case where the delay is assumed deterministic and equal to the current mean with the base case results in a less dramatic difference, of course: the number of ambulances needed increases from 15 to 16. However, the impact of ignoring the variability in delays would be far greater if the mean delay were higher. For example, if the mean delay were to increase by 25% (from 2.6 minutes to 3.25 minutes), while the standard deviation stayed the same, then 21 ambulances would be needed to reach the coverage goal. In this case, if the delay variability were ignored (i.e., the standard deviation is assumed to be zero), then the model predicts that only 18 ambulances would be needed to reach the coverage goal. Hence, a model that incorporates delays but treats them as

deterministic would underestimate the number of ambulances needed to provide the target coverage by (21-18)/21 = 14%.

Figure 3-6 gives the complementary perspective and provides additional insight into the impact of the delay standard deviation. It demonstrates how the system wide coverage varies when the parameters of the delay distribution are varied in the same way as for the results in Figure 3-5, with the total number of ambulances fixed at 16. From Figure 3-6, it is apparent that if the variability in the delay is not considered, then the estimated coverage is about 93%, compared to just over 90% if the variability in the delay is incorporated. When the standard deviation of the delay is decreased 50% from the base case, the coverage increases from just over 90% to about 92%. When the standard deviation is increased 25% from the base case, the coverage drops to about 89%. The results are magnified as the average level of the delay increases. These results illustrate the importance of accounting for delays in order to obtain accurate estimates of the coverage and of the resources required to attain a specified coverage. They also illustrate the importance of controlling the call-taking and dispatching processes to ensure that delays do not increase (but preferably, decrease).

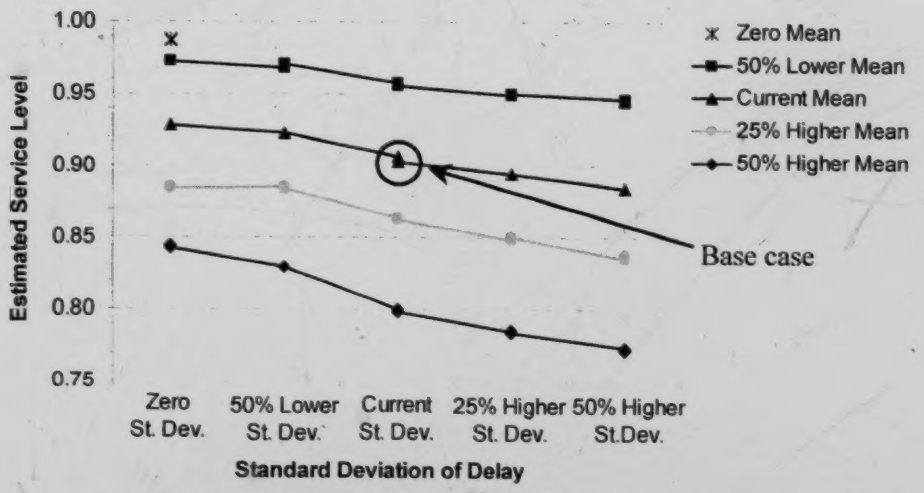


Figure 3-6: Sensitivity of the system wide coverage to the mean and standard deviation of the delay distribution, when the total number of ambulances is fixed at 16.

Discussion

This section outlines several possible avenues for further research involving exploration of the optimization model (P1), its properties, solution approaches, and insights from its application. First, three extensions of the model that are fairly straightforward are discussed, and then some avenues for further research are examined.

Model Extensions

One can add a constraint to (P1) to ensure that the probability that at least one ambulance is available is above some threshold β, as follows (assuming independence between ambulances, as in (6)):

$$1-\prod_{j=1}^{J}\rho_j^{x_j} \geq \beta \tag{8}$$

The constraint can be linearized by isolating the product of the busy fractions on one side of the inequality and taking logarithms of both sides, resulting in:

$$\sum_{j=1}^{J}\ln(\rho_j)x_j \geq \ln(1-\beta) \tag{9}$$

Note that preliminary experiments using data from Edmonton indicated that the coverage constraint (3) was tighter than constraint (9) for values of $\beta \geq 0.99$.

In addition to the constraint (3) on the system-wide coverage, one could add constraints on the coverage for each demand node, of the form

$$c_m(\mathbf{x}) \geq \alpha_m, \text{ for } m = 1,2,...,M \tag{10}$$

where α_m is the target coverage for demand node m. This constraint set could, for example, be used to impose a common minimum coverage for all demand nodes or some subset of the demand nodes.

One can also add variables and constraints to decide which stations to open and to limit the number of ambulances at each station. Specifically, let y_j be a binary indicator variable for whether station j is opened; let a_j be the fixed cost of opening station j; let d_j be the variable cost of locating one ambulance at station j; and let b_j be the maximum number of ambulances at station j, if it is opened (if there are no such limits, then one can set $b_j = B$ for some sufficiently large number B). The extended problem formulation is:

(P2) minimize $\sum_{j=1}^{J}(a_j y_j + d_j x_j)$

subject to (3), (4), (8), (9)

$x_j \le b_j y_j$, for $j = 1, 2, \ldots J$

$y_j \in \{0,1\}$, for $j = 1, 2, \ldots J$ (11)

Note that the continuous relaxation of (P2) is a convex programming problem. However, (P2) is more difficult to solve than (P1) because it has more integer variables.

Further Research

Incorporation of random delays and travel times may influence not only the total number of ambulances needed to provide a given level of service, but also how ambulances are distributed through the system. Experiments to generate insight into whether this happens and how are planned. In order to do further computational testing of the model, data from a city of similar size to Edmonton, but which is aggregated into many more (smaller) zones and has up to 40

potential locations for ambulances will be used. We also hope to use the model to estimate the impact of various changes to the operation of an ambulance system. For example, it may be possible to reduce delays by performing activities in parallel rather than in series, but such a change may increase ambulance workload, if it results in more false alarms. Therefore, we would like to explore the trade-off between reducing delays and increasing busy fractions. Estimation of the travel time distribution functions $H_{jm}(t)$ is likely to be challenging. The chapter that follows focuses on models for estimating the travel time distribution.

Conclusions

An optimization model for allocating a minimum total number of ambulances to stations so as to satisfy a system-wide coverage constraint was presented. The model differs from previous related work in that the variation in pre-travel delay is considered (in addition to the variation in travel time) when calculating the fraction of demand that is covered within the time standard. Data from recent projects with the town of St. Albert and the City of Edmonton indicate that pre-travel delays are important and highly variable (with a standard deviation of about 40% of the mean). Computational experiments demonstrate that the inclusion of the variability of such delays has a substantial impact on the solution that the model prescribes. The formulation is sufficiently tractable that it can be solved to global optimality for cities with population around one million aggregated to around 180 demand nodes, and with as many as 10 stations with general-purpose solvers.

References

R. Batta, J. Dolan, and N. Krishnamurthy (1989). The Maximal Expected Covering Location Problem: Revisited. *Transportation Science* **23** 277–287.

O. Berman and D. Krass (2001). Facility Location Problems with Stochastic Demands and Congestion. In *Location Analysis: Applications and Theory*, eds. Z. Drezner and H.W. Hamacher. Springer Verlag.

M. Brandeau, and R.C. Larson (1986). Extending and Applying the Hypercube Model to Deploy Ambulances in Boston *Deliyery of Urban Services*, eds. A. Swersey and E. Ignall. North Holland, New York.

R. Church and C. ReVelle (1974). The Maximal Covering Location Problem. *Papers of the Regional Science Association* **32** 101-120.

M.S. Daskin (1983). A Maximum Expected Covering Location Model: Formulation, Properties, and Heuristic Solution. *Transportation Science* **17** 48–70.

M.S. Daskin (1987). Location, Dispatching, and Routing Model for Emergency Services with Stochastic Travel Times. A. Ghosh, G. Rushton, eds. Spatial Analysis and Location-Allocation Models. Van Nostrang Reinhold Company, New York, 224–265.

D. J. Eaton, M.S. Daskin, D. Simmons, B. Bulloch, and G. Jansma (1985). Determining Emergency Medical Service Vehicle Deployment in Austin, Texas. *Interfaces* **15** 96-108.

Frontline Systems, Inc., Premium Solver, www.solver.com/xlspremsolv.htm/.

J. Goldberg, R. Dietrich, J. M. Chen, M. G. Mitwasi, T. Valenzuela, and E. Criss (1990a). A Simulation Model for Evaluating a Set of Emergency Vehicle Base Locations: Development, Validation, and Usage. *Socio-Economic Planning Sciences*, **24** 125-141.

J. Goldberg, R. Dietrich, J. M. Chen, M. G. Mitwasi, T. Valenzuela, and E. Criss (1990b). Validating and Applying a Model for Locating Emergency Medical Vehicles in Tucson, AZ. *European Journal of Operational Research* **49** 308–324.

J. Goldberg and L. Paz (1991). Locating Emergency Vehicle Bases when Service Time Depends on Call Location. *Transportation Science* **25** 264–280.

L. Green and P. Kolesar (1989). Testing the Validity of a Queueing Model of Police Patrol. *Management Science* **35** 127–148.

S. G. Henderson and A.J. Mason (2000). Development of a Simulation and Data Visualisation tool to assist in Strategic Operations Management in Emergency Services. School of Engineering Technical Report 595, University of Auckland, New Zealand.

S. G. Henderson, and A. J. Mason (2004). Ambulance Service Planning: Simulation and Data Visualisation. To appear in *Handbook of OR/MS Applications in Healthcare*, eds. F. Sainfort, M. Brandeau, and W. Pierskalla, Kluwer.

A. Ingolfsson, E. Erkut, and S. Budge (2003). Simulating a Single Start Station for Edmonton EMS. *Journal of the Operational Research Society*, **54** 736-746.

J. Jarvis (1985). Approximating the Equilibrium Behavior of Multi-Server Loss Systems. *Management Science* **31** 235–239.

R.C. Larson (1974). A Hypercube Queueing Model for Facility Location and Redistricting in Urban Emergency Services. *Computers and Operations Research* **1** 67-95.

R.C. Larson (1975). Approximating the Performance of Urban Emergency Service Systems. *Operations Research* **23** 845-868.

R.C. Larson (1979). Structural System Models for Locational Decisions: An Example Using the Hypercube Queueing Model. *Operational Research '78, Proceedings of the Eighth IFORS International Conference on Operations*

Research, ed. K. B. Haley. North-Holland Publishing Co., Amsterdam, Holland.

V. Marianov and C. ReVelle (1995). Siting Emergency Services. *Facility Location: A Survey of Applications and Methods*, ed. Z. Drezner, Springer.

V. Marianov and C. ReVelle (1996). The Queueing Maximal Availability Location Problem: A Model for the Siting of Emergency Vehicles. *European Journal of Operational Research* **93** 110–120.

C. ReVelle and K. Hogan (1988). A Reliability-Constrained Siting Model with Local Estimates of Busy Fractions. *Environment and Planning B: Planning and Design* **15** 143–152.

C. ReVelle and K. Hogan (1989). The Maximum Availability Location Problem. *Transportation Science* **23** 192–200.

C. Saydam and H. Aytug (2003). Accurate Estimation of Expected Coverage: Revisited. To appear in *Socio-Economic Planning Sciences.*

A.J. Swersey (1994). The Deployment of Police, Fire, and Emergency Medical Units. *Handbooks in Operations Research and Management Science, Vol. 6: Operations Research and the Public Sector*, eds. S.M. Pollock, M.H. Rothkopf, A. Barnett, North-Holland.

C. Toregas, R. Swain, C. ReVelle, and L. Bergman (1971). The Location of Emergency Service Facilities. *Operations Research* **19** 1363-1373.

T. R. Willemain and R. C. Larson, eds. (1977). *Emergency Medical Systems Analysis*. Lexington Books, Lexington, MA.

Chapter 4: Empirical Analysis of Ambulance Travel Times

Introduction

An important component of virtually any model of emergency service operations, whether an analytic model, a simulation model, or an optimization model, is the travel time of the vehicles, especially of those vehicles enroute to a call. This is true because often the most critical performance measure for such a system is the time it takes to respond to an emergency, and the travel time of the vehicle to the scene of the emergency is typically a large portion of the response time. On the surface, the problem of estimating travel times may seem trivial, one might think it is simple – just divide the travel distance by the speed. However, if we look beyond the surface, the problem is anything but trivial. First, what should be used for the travel distance? Given location information (possibly x and y coordinates, or latitude and longitude) for an origin (say an ambulance station) and a destination (say an emergency scene), a number of methods can be used to estimate the distance between the two. Second, what should be used for the travel speed? Perhaps an average historical speed between the two points, or maybe a weighted average of the speed limits on the roads between the two points would be useful. Should the speed be dependent on factors such as the time of day, day of week, month of year, weather, types of roads between the two points, or type of vehicle? Next, is it necessary to consider acceleration and deceleration and if so, at what level of detail? Finally, what about randomness in the travel times? Even for the same origin and destination, the travel time will have some variation due to different routes, different drivers, different traffic conditions, or as a result of the aggregation of points into zones, or other known and unknown factors. It quickly becomes apparent that the problem is much more complicated than it first appears.

In this chapter, the problem of estimating travel times for emergency service operations is considered in detail. Various data from several sources are employed throughout the chapter. These datasets allow the specifics of travel time in ambulance operations to be examined and some of the assumptions and methods that have been used in the literature to estimate and model travel time distributions to be considered. The specific types of data used are quite unique and provide a rare opportunity for this type of research.

The remainder of this chapter is structured as follows. In the next section, a review of the literature on estimating travel times for emergency service vehicles is given. Following that is a discussion of the data: what it includes, how it is collected, and some of its limitations. Then the methodology used is outlined and the results of the analysis are reported. The chapter closes with conclusions and some comments on areas for further research.

Literature Review

A good overview of the problem and review of the early literature is given by Walker, Chaiken, and Ignall, (1979) in the context of fire department deployment. In this section, the concentration is on those papers that are most relevant to the research undertaken in this chapter. The focus is on models where the origin and destination are known, except for possible aggregation and measurement errors. For this reason coverage of models for estimating travel times in the case where there is complete uncertainty about locations is omitted, even though there are such models that have been developed with a focus specifically on emergency services (see for example Kolesar and Blum, 1973, Kolesar, 1975, Walker, Chaiken, and Ignall, 1979, and Larson and Odoni, 1981).

The discussion here is organized into three subsections. The first of the subsections will focus on estimation of travel distances, the next on estimation of expected travel times, and the final on random variability in travel times.

Travel Distances

There are essentially two different approaches for estimating travel distances between a given origin and a given destination to be used within models for emergency service systems.

The first approach is to use a metric based on the coordinates (x_i, y_i) of the origin and the coordinates (x_j, y_j) of the destination to estimate the "point-to-point" distance d_{ij}. These metrics range from the Euclidean distance, to rectilinear distance, to more general distance functions as given below.

- rectilinear (right angle)

$$d_{ij} = |x_i - x_j| + |y_i - y_j|$$

- Euclidean (straight line)

$$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$$

- modified Euclidean

$$d_{ij} = k\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$$

- general function

$$d_{ij} = k\left[|x_i - x_j|^p + |y_i - y_j|^p\right]^{1/s}$$

Love and Morris (1979) evaluated these (rectilinear, straight line, and more general) functions and provide some accuracy results. For the modified Euclidean equation they gave best-fit values from 1.16 to 1.28 for the constant k for five cities in their sample. Note that although the general functions can provide more

accurate results compared to the rectilinear or Euclidean functions (Love and Morris, 1979; Brimberg and Love, 1992), actual distance data are required in order to estimate the parameters for a particular region. The perhaps not so obvious catch, especially for travel within a city is that the "actual distance" may not be known (or even deterministic) since it will depend on the route taken. An extension to this method that involves the calculation of additional distance for barriers to rectilinear road travel is discussed in a paper applying the hypercube queueing model to deploy ambulances in Boston (Brandeau and Larson, 1986).

A second approach for estimating distances is to use a network model with distances for each link and calculate the shortest path between the origin and the destination. Unlike the first approach, generally in a network model the origin and destination are not specific points, but rather the points have been aggregated into zones. The network data are typically stored in a geographic information system (GIS) and such systems can potentially contain various levels of aggregation. However, even though it is possible to have a very low level of aggregation in the network distances, a higher level of aggregation may be necessary in order to work with a shortest path algorithm. The other major difference is that the actual road network is captured, so that travel distances will be based on actual road travel distances with the only errors being due to aggregation and uncertainty in route choice. Uster and Love (2003) suggest that distance-predicting functions (such as those given above) can be preferable to storing (and working with) large files of network distance data and provide a method for calculating confidence intervals for the predicted distances. In order to calculate confidence intervals they use a sample of origin-destination pairs with "actual" distances, and thus incorporate such sources of uncertainty as barriers to travel in the road network that are not accounted for in the distance estimates, measurement errors in the point coordinates and "inaccurate instrument calibrations".

Travel Times

The approaches for estimating travel times mirror those of estimating travel distances with a few extra complications.

The first approach considers point-to-point travel times and relates the point-to-point distance to the travel time. One method that falls under this first approach, as described in Walker, Chaiken, and Ignall, (1979) and attributed to Hausner, uses two different relationships depending on the length of the trip. For short trips time is assumed to increase with the square root of the distance, while for long trips it is assumed to increase linearly with distance.

$$\mathrm{E}\left[T_{ij} \mid D_{ij}\right] = \begin{cases} c\sqrt{D_{ij}} & D_{ij} \le d \\ a + bD_{ij} & D_{ij} > d \end{cases}$$

The parameter values (a, b, c, and d) are estimated from data for a city or region, although Walker, Chaiken, and Ignall, (1979) claim that there is little variation in these numbers between cities.

Kolesar, Walker, and Hausner, (1975) provide a set of assumptions for a trip between some origin and destination, that imply this same functional form and give meaning to the parameters. The assumptions are that beginning at the origin, the unit accelerates at a constant rate, a, to a constant cruising velocity, v_c, and then travels for some time at that velocity and finally decelerates (at the same constant rate, a) to stop at the destination. Note that d_c is the distance required to achieve the cruising velocity and is a function of a and v_c rather than an independent parameter. The derivation for this relation is given in Appendix 2 (appendices are located at the end of this dissertation), and the details for the method that was proposed by Kolesar, Walker, and Hausner, to fit this function to experimental data are provided in Appendix 3.

$$E\left[T_{ij} \mid D_{ij}\right] = \begin{cases} 2\sqrt{D_{ij}/a} & D_{ij} \le 2d_c \\ v_c/a + D_{ij}/v_c & D_{ij} > 2d_c \end{cases}$$

From this point forward, this relation will be referred to as the Kolesar-Walker-Hausner (or KWH) function. Figure 4-1 gives a pictorial view of the speed time profile that is assumed for a given trip between two points. Here, t_c is the time required to achieve cruising velocity and T is the total length of the trip. In this picture, it is assumed that the trip is long enough for the vehicle to reach the cruising speed (i.e., $T > 2t_c$).

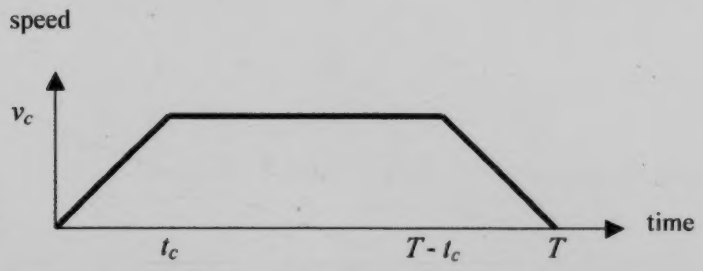


Figure 4 - 1: Speed time profile assumed for the KWH function.

In reality, an emergency service vehicle will not travel in exactly this way (constant velocity preceded by constant acceleration and followed by constant deceleration), rather the speed of the unit will vary in a much more complicated manner with possible stops at red lights and stop signs, possible slowdowns due to traffic or weather, and different speeds on different road types (residential roads versus main arteries). However, this model may still provide a good approximation to the travel time for the vehicle travelling between a given origin and destination. In a field experiment with fire response vehicles in New York City, Kolesar, Walker, and Hausner, (1975) found that the travel times for regions with shorter response distances tended to follow a square root relation and for regions with longer distances tended to follow a linear relationship. Further, they found that although there were statistically significant differences between the

parameters for different companies within each of the two region types, the differences were not large and a good fit to the average travel times for all regions of the city was found using the combined relation above. Additionally, they reported that the hour of the day had little effect on the average travel velocities of the fire vehicles (rush hour speeds were about 20% lower than non-rush).

Carson and Batta (1990) state that not modelling the "elbow" relationship (which occurs at $D_{ij} = 2\ d_c$ in the KWH function) between travel time and distance can result in serious errors in predictions of system performance. They found that the predicted savings of 30% in system-wide average response time for their proposed operational changes for a one ambulance system (on a university campus) was reduced to 6% in test runs and they attributed this to the fact that they assumed a constant average speed regardless of travel distance.

A related method that falls under the scope of this first approach involves using linear regression to describe the relation between the point-to-point distances and the travel times, in some cases incorporating additional factors, such as road types, seasonal or time of day variations, or type of trip, into the relationship. Without the additional factors the model is comparable to the previous models for longer distances. If D_{ijk} is the distance along roads of type k and the X_l's are the additional factors considered then the model would be as follows.

$$\mathrm{E}\left[T_{ij} \mid D_{ijk}, X_l\right] = a + \sum_k b_k D_{ijk} + \sum_l c_l X_l$$

The second approach involves using road network data and calculating fastest (as opposed to shortest) paths. The reasoning behind this approach (as compared to the network approach in the previous section) is that the paramedics, seeking to get to the site of the emergency as quickly as possible, might choose routes that are longer if they expect that the travel time will be shorter due to higher average speeds along those routes. Note, however, that it is possible that the actual route

taken for a particular emergency is not the route that would be expected using either the shortest distance or fastest time network calculations. The regression approach mentioned above can also be used in concert with this method, and in this case generally incorporates travel distance on a number of road types as explanatory variables. For example, Erkut et al. (2001) apply an algorithm to compute fastest paths on a network in concert with a linear regression method incorporating 3 road types, time of day (rush vs. non-rush) and season (winter vs. summer) as determinants of travel speeds. Similarly, Goldberg et al. (1990) used a network model to calculate distances (along a planned route) between each station and demand node and used linear regression to determine the travel speeds on 4 different road types. An additional factor that could potentially have an impact on the route taken by an emergency service vehicle, the time of the day, is considered by Henderson and Mason (2000, 2004) using a network approach. They use time varying travel times and compute time dependent fastest paths (heuristically) for use in a simulation model.

Cook and Russell (1980) compare point-to-point travel time estimates for Tulsa, Oklahoma of three methods: an extension of the KWH function, linear regression (with distance only and with distance and weighted average speed limit between the points, as independent variables) and a software package that uses a network approach (with 448 population centroids). For the first two methods a rectangular distance metric was used. The authors found that the network approach was not as accurate as the other two methods (perhaps due to aggregation in the network data used), which were comparable to each other in accuracy. In addition, they found that adding the average speed limit between the two points as an independent variable in the linear regression method provided little additional explanatory power.

Randomness in Travel Times

A number of researchers have considered randomness in travel times in various emergency service problem contexts, under differing assumptions and using different methods for modelling the randomness. For the problem of estimating the travel time given an origin and destination, the following are potential sources of variation;

1. Variation that can be explained using readily available data, for example time of day, day of week, and season.
2. Variation that could perhaps be explained, if more data were available, for example the precise location of the responding ambulance (whether at a station or other location, idle, or near a station, already moving), the precise location of the call, weather, traffic light locations and cycles, and so on.
3. Variation due to factors that are either difficult to obtain data on or unknown. Factors that are difficult to obtain data on but seem likely to impact travel time include route choice, driver behaviour, and unpredictable and extreme weather and traffic events.

The discussion of the literature in this section is organized around the primary purpose of the travel time variability in each paper. The assumptions made about the source of the variability and the approaches taken to model the variability are discussed.

Some of the research in this area is not directly focused on modelling the travel time distribution, but rather on how the incorporation of travel time variability can affect locational or other operational decisions. For example, Daskin (1987) uses a normal distribution to incorporate travel time variability in his multi-objective model for determining vehicle locations/allocations as well as the dispatch policy and routes that the vehicles should take. He assumes that travel times on "non-

overlapping links" are independent, and that the variance of the travel time on each link is proportional to the mean travel time of the link (the proportionality constant is the same for all links). With regard to the assumption that the link travel times are normally distributed, he cites Daskin and Haghani, 1984 who compare this assumption to the more realistic (in the sense that negative travel times are not allowed and the distribution of travel times is not assumed to be symmetric about its mean), but also more computationally burdensome, Erlang distribution and find it is reasonable. Similarly, Mirchandani and Odoni (1979) consider the p-median problem with discrete random variable travel times and find that travel time variability can significantly affect the location decisions. They use a network approach and emphasize that using expected values in place of the probability distributions for the link travel times in order to calculate fastest paths can lead to invalid results. Given that they are using a network approach, they consider variations (and especially predictable variations) in traffic to be the source of the randomness in travel times. They break the travel times down into individual link travel times allowing the travel time along each link to follow an arbitrary discrete distribution and they allow the distributions for two separate links to be dependent.

Other researchers have made specific assumptions about the source of the variability and direct attempts to develop models for the variability (i.e., for the probability distribution of the travel time) based on those assumptions. Aly and White (1978) develop a probabilistic formulation of the emergency service location problem under the assumption that the stochastic variation is due to randomness in the locations of incidents. They determine the probability distribution for the travel time between the facility locations and the random location of the incident under the assumptions of constant travel speed in a given area and potential incident locations being uniformly distributed over a rectangular area. This method could provide a useful way to deal with demand aggregation. Another paper that has a direct focus in this area uses the hypercube

model (Larson, 1974) to estimate the travel time distribution in a particular region (Chelst and Jarvis 1979). In that paper the authors assume that the travel time between two nodes in a network is known and deterministic, and consider the randomness to arise from the variation in the location of the responding unit (due to a difference in the dispatch order for different locations within the region, or as a result of ambulance unavailability). They note that average travel times within a region are dependent on the deployment pattern within the region and measures that do not account for this will suffer in accuracy.

Finally, a number of authors have incorporated travel time variability into their models for particular case studies in order to provide more realistic estimates of system performance. Goldberg et al. (1990) develop a travel time model to estimate a distribution for the travel time between each base-zone pair using a network type approach. They use the (average) data from actual calls to fit a regression model (with "actual" distances on four road types as explanatory variables) to estimate the mean travel times, and then use the differences between these predicted averages and the actual call travel times to get a distribution of the deviation from the predicted travel time mean. This implicitly assumes that the travel time variance is constant and does not depend on the distance. In their model, they determined the "planned route" for each pair and then measured the distance. All vehicles were assumed to start from their home base and actual travel times were recorded to the nearest minute.

What seems to be lacking in the literature is an examination of the distribution (and in particular the variability) of actual travel times for emergency service vehicles and a study on how to incorporate that distribution into methods for estimating the travel time as a function of distance. That is the focus of the remainder of this chapter.

Data

Format of the Data

We have used data from three cities in the investigation of travel times and variability in travel times of emergency response vehicles. First, we have data from Edmonton EMS including one day's worth of automatic vehicle locator (AVL) data, event (transaction) data for the same day, as well as event data for one year for a particular demand location. We also have event data (for three full years and portions of two additional years) from Calgary EMS and (for one month in 1999) from Montreal.

The AVL data include location information (latitude and longitude) for each on-duty ambulance in the city, along with vehicle status and event numbers. The information is collected every minute for stationary units and about every 150 metres for vehicles in motion. These data allow an in depth look at the movements of the vehicles as they travel throughout the city responding to calls, transporting patients to hospitals, and returning to stations to await calls. To our knowledge no one has used AVL data to perform the types of analysis reported here.

The event data consist of timestamps generated by humans for the various epochs in the process of ambulance service for an emergency call, and also include location information (for the call, the ambulance at time of dispatch, and the hospital that the patient is transported to). The event data allow an investigation of the distribution of travel times between particular origin destination pairs as well as an examination of various methods of estimating travel times of emergency response vehicles.

Data Issues

The data employed in this analysis include errors from a number of sources. The integrity of the event data depends on the ability of the paramedics, to first recall to indicate their status at the exact moments that there is a change in their status, and then to do so without error. Keeping in mind that a primary responsibility of the emergency service personnel is to respond as rapidly as possible to calls for service in order to reduce the suffering of persons in emergency situations, certain errors in the data are to be expected. In fact upon examination of the data it is clear that there are records where the data are not accurate due to disruptions in this process. For example, there are instances when the travel time to the scene is unexpectedly long (such as 30 minutes) and the time spent on scene with the patient is unexpectedly short (10 seconds). The obvious explanation is that the timestamp for the arrival of the ambulance at the scene was not recorded correctly; rather, the change in status (from "enroute" to "arrival at scene") was not indicated until the next change of status ("departure from scene") occurred. Clear errors such as this can be removed from the data, but there may be other errors that are not as easy to detect.

Although the AVL data do not suffer from this type of distortion (with the exception of the unit status indicator field), there can also be errors in the AVL data that could affect the results of the analysis. There are numerous potential sources for error in this type of data, from satellite clock errors, to signal delay errors as a result of the earth's atmosphere, to "selective availability," a term for the intentional degradation of the accuracy of the satellite signals for non-military users prior to May 2000. Detailed discussions of these and other sources of error and how to deal with them can be found in the guide of Geomatics Canada (1993) or the text by Strang and Borre (1997). For an example of how such errors can affect the AVL data, consider Figure 4-2 created from the Edmonton AVL data. The graphs show speed vs. time profiles for two particular (emergency) events.

The black and dark grey lines on the graphs are the average speeds (in km/hr) of the ambulances, as calculated using right angle and straight line distance metrics, respectively, from the location and time information in the AVL data, while the light grey line is the speed (in km/hr) recorded in the AVL data. The points where the status of the responding unit changes appear along the x-axis. In the first graph there is a spike that indicates an obvious error in that the calculated speeds exceed 200 km/hr while the speed recorded by the AVL is a more credible 50 km/hr. Although access was not obtained to information about how the speed is calculated in the AVL data, it is assumed that it is either an instantaneous speed calculated using a method that does not rely on the position information (such as the Doppler shift method mentioned in the text by Strang and Borre, 1997), or that a smoothing or filtering algorithm has been used to reduce the impact of errors in the position and time measurements. In either case, it appears to be a more reliable estimate of the vehicle speed (compared to the values obtained using the position and time information from the AVL data) and so any subsequent discussions involving speed in the AVL data will focus on the speed recorded by the AVL system. This graph also shows that the unit is classified as enroute to the call a substantial amount of time (a couple of minutes) before it actually starts moving. The second graph shows a number of spikes and discrepancies between the calculated speeds and the speeds recorded by the AVL as well as a considerable (about 12 minute) delay between the time the ambulance stops moving (when classified as transporting a patient) and the time the status changes to "transport arrived". In this regard, it can be seen that the AVL data allow one to assess the accuracy of some of the timestamps in the corresponding event data.

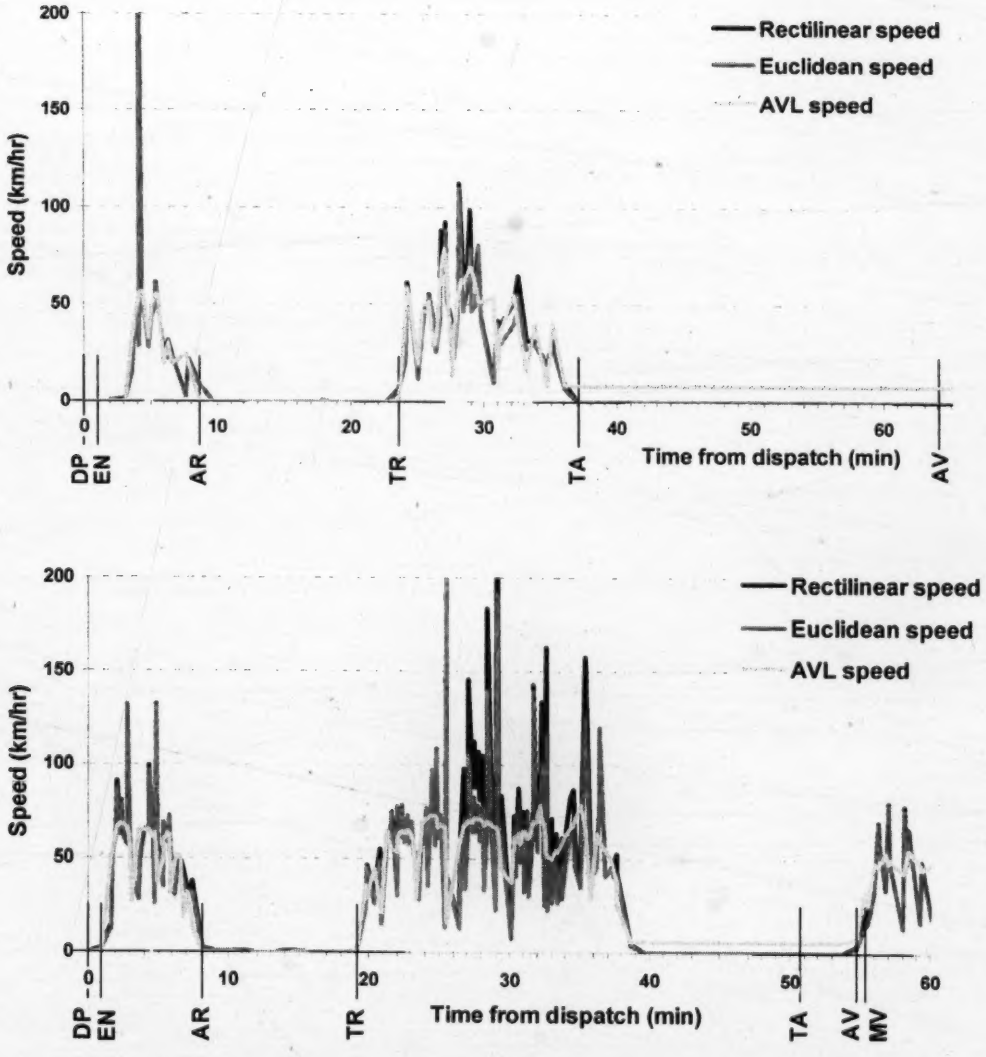


Figure 4 - 2: Speed time profiles of ambulances responding to events. The unit status appears along the x axis where DP stands for "dispatched", EN for "enroute", AR for "arrived on scene", TR for "transporting patient to hospital", TA for "transport arrived at hospital", AV for "available" and MV for "moving (returning to a station)".

In the next subsection the event data are examined closely to see what types of outliers and errors are observable and potential candidates for removal. Also, the type of scrutiny just discussed in this subsection and the type that will be

discussed in the next subsection could be valuable for identifying potential modifications to the data collection process to reduce the number of errors. For example a simple modification would be to have two buttons in the ambulance: one for "catch-up" and one for "timestamp." Then, if the paramedics forgot to press the "timestamp" button when arriving on the scene of a major accident, when they are ready to transport a patient, they could then press "catch-up" first (to indicate that they forgot to indicate when they arrived on scene) and "timestamp" second, to indicate that the "begin transport" timestamp is accurate. The paramedics themselves can probably think of other simple and useful modifications to the process.

Outlier Analysis

A number of steps were taken to attempt to ensure the integrity of the data before using it in the analysis. To this end a number of different options were considered and an overview of them is given next, followed by more discussion in the results section of the chapter. One concern was that data that were potentially meaningful not be removed simply because they looked like they did not fit with the rest of the data. In particular, because of the nature of the variable under consideration, it would not make sense to consider a very large (or even a very small) travel time an outlier on its own, without reference to some other indication of inaccuracy. One thing that could signal a potential large outlier in terms of travel time actually being the result of an error (that should be removed) is a corresponding small outlier in one of the other time components and vice versa. So one of the methods used to clean the data was to remove records that had an "extreme outlier" of opposing magnitude in a field next to the travel time (either the pre-travel delay time or the on-scene time), or in any of the travel time interval or distance fields. Values greater than three times the interquartile range (the range of the middle half of the data) from the first or third quartiles (the boundaries for the interquartile range) were considered to indicate extreme

outliers. Due to heavy skewness in many of the intervals in the datasets, none of the lowest values came up as outliers for any of the fields. As a result, a log transformation was applied to the time intervals closest to the travel time in order to provide more symmetric distributions and records that subsequently indicated significant outliers were removed. As a final method for identifying erroneous data, average speeds were calculated for each record using both Euclidean and rectilinear distance metrics from the locational information in the datasets, and those records with speeds below 10 km/hr at the low end and above 100 km/hr at the high end were removed.

Figure 4-3 shows box-plots to illustrate the distributions for the travel times (for the highest priority calls in the year 2003 of the Calgary data) with various records removed as outliers. Note that for this figure as well as for Figure 4-4, there was no filtering done on the travel distance, so the distributions include travel times aggregated over all distances. The first box plot is for the original data, with no records removed. The second has all records with missing values for any of the time interval fields removed. There were two reasons that it was deemed necessary to do this. First, missing values in the time intervals next to the travel time could indicate that that value was amalgamated with the travel time. Second, missing values for the stages following the travel time could indicate that travel to the scene was not completed, or that the patient was not transported to the hospital, which could in turn signify some difference in the type of call that could have had an impact on the response time. The last four box plots all represent different methods of removing potentially erroneous data from the second dataset. The third and fourth sets result from removing records that were outliers in the original distributions, and the log transformed distributions respectively, of the time interval fields closest to the travel time field. The fifth and sixth sets result from removing records that have speeds less than 10 km/hr or greater than 100 km/hr, based on Euclidean and rectilinear distance metrics respectively. As would be expected, since the distributions are so skewed and no

outliers are removed at the low ends of the time intervals surrounding the travel time from set 2 to set 3, the mean of the travel time distribution will increase (removing outliers from the high ends of the surrounding time intervals will tend to remove the low "errors" from the travel time distribution, with no corresponding adjustment to the high end. The result is an even more skewed distribution as evidenced by the box plot for set 3. In contrast, the box plots for sets 4 through 6 show less skewness than set 2. This suggests that a better way to remove erroneous data would be to transform the data to a more symmetric distribution first (as in set 4) or to consider setting upper and lower limits on the average travel speed (sets 5 and 6). A final note on the results here is that depending on the method used to remove outliers, the mean (as well as the standard deviation which is not shown in the figure) can vary to a large degree, while the median and interquartile range are relatively stable.

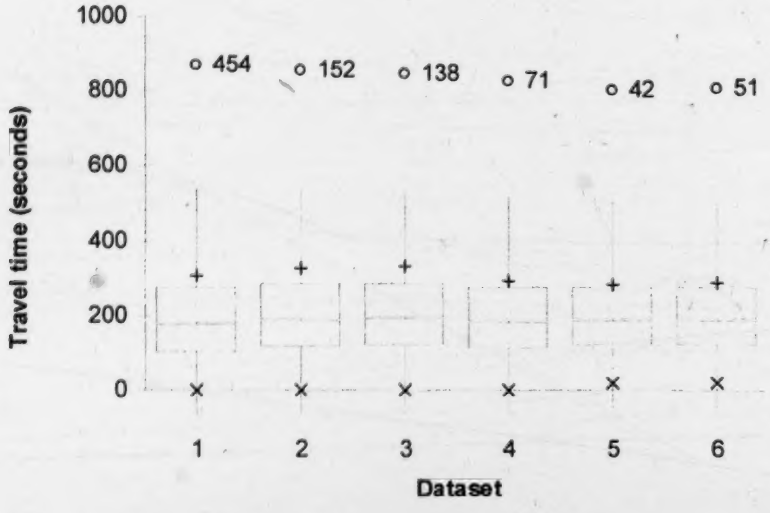


Figure 4 - 3: Box plots of the travel times for various datasets: "+" indicates the position of the mean, "x" indicates the position of the minimum or maximum observation in the event that it is within three times the interquartile range, "o" indicates outliers beyond three times the interquartile range, and labels next to the "o's" show the number of outliers beyond that range.

With the exception of the outliers at the higher travel times and the positioning of the means, it may be difficult to see the overall skewed shape of the distributions from the box plots, so the probability density function (PDF) and cumulative distribution function (CDF) for dataset 6 are shown in Figure 4-4 along with the corresponding fitted lognormal distributions. The choice of the lognormal distribution will be discussed in a later section.

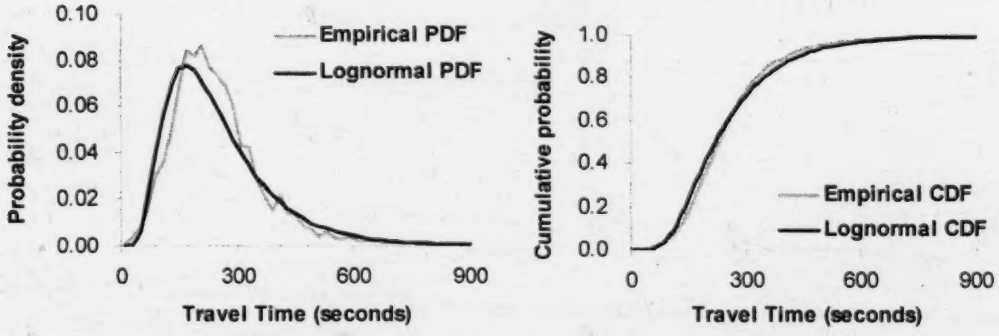


Figure 4 - 4: Probability density and cumulative distribution functions for Dataset 6.

In examining the data it is clear that there are outliers that are obviously due to errors as well as some other outliers that are likely due to errors and that many of these could have an impact on estimation procedures. So in the analyses completed, this was kept in mind and an attempt was made to focus on estimators and estimation methods that would be more robust to such issues. Most of the subsequent analyses show results using the original (uncleaned) data but in some cases outliers were removed, typically using the method of dataset 4 discussed above (removing records with outliers in the log transformed fields next to the travel time). In cases where data have been removed, it is indicated in the text.

Methodology

The research on travel times and travel time variability can be separated into two stages. The first stage consists of preliminary analysis of the data. The second stage involves a more in depth examination of various methods for estimating emergency vehicle expected travel times and modelling travel time variability.

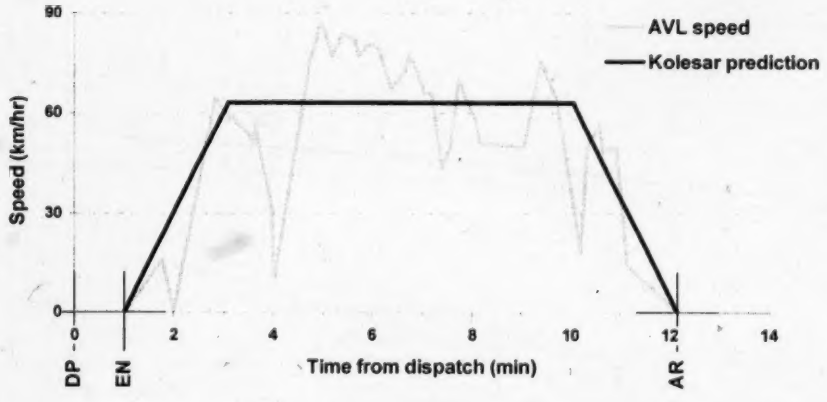
The preliminary analysis phase of the research consists of a detailed examination of the data in order to gain insight into the individual facets of the ambulance movements in relation to a call for service. In particular, the travel time distribution for a particular station/demand location pair is examined to get a sense of the variability in the travel time that is not caused by variations in the location of the responding vehicle or the call. In addition, speed time profiles created using the AVL data are scrutinized to focus on the details of the ambulance travel.

In the second phase of the study the event data are used in order to compare different methods for estimating travel times of emergency vehicles. Much of the discussion here will focus on the KWH travel time function. Of particular interest is how to best make use of the function in conjunction with the many potential methods for estimating travel distance, and what other factors, if any, should be incorporated. In the original paper Kolesar, Walker, and Hausner, (1975) used a field experiment in New York City to calculate actual travel times and distances (by odometer reading) for a number of runs of fire companies (ranging from 12-150 for an individual company) in order to estimate the relationship. For this field experiment they found the following values for the parameters; acceleration cutoff distance, d_c, of 0.44 miles (0.7 km), average cruising speed, v_c, of 39.2 mi/hr (63.1 km/hr), and acceleration, a, of 29 mi/hr/min (47 km/hr/min). In addition to focusing on estimating expected travel times, consideration will be given to modelling the variability in the travel time distribution.

Results

Preliminary Analysis

We begin by exploring the data to see what types of patterns and relationships we can observe. Figure 4-5 shows speed vs. time profiles for three events, similar to those in an earlier section, but with a focus on the "travel to scene" component of the service. Despite the variability in the travel speed over the trip, the KWH travel time function would appear to provide a relatively good approximation to the behaviour of the ambulances responding to an emergency call. The parameters for the piecewise linear curve that was visually fit to the speed profiles are: cut-off distance of 2.2 km, cruising velocity of 63.0 km/hr, and acceleration of 30.0 km/hr/min. The velocity is quite close to the values reported for fire departments by Kolesar, Walker, and Hausner, (1975) and Larson and Odoni (1981), but the acceleration is lower and the cut-off distance is higher.



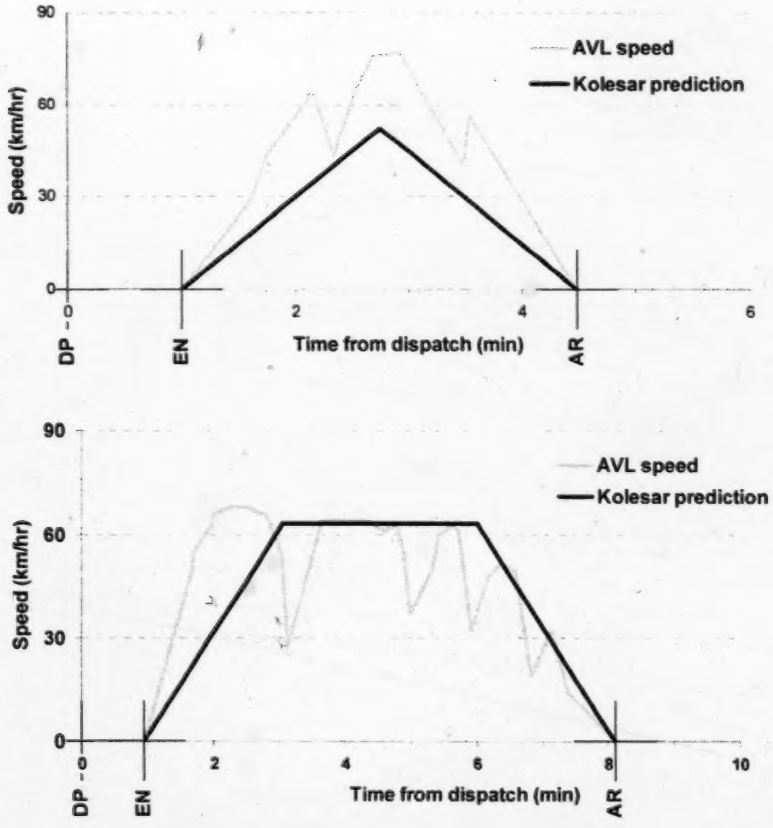


Figure 4 - 5: Speed time profiles of the travel to scene component of ambulances responding to events. The unit status appears along the x-axis where DP stands for "dispatched", EN for "enroute", and AR for "arrived on scene".

Next, factors that might have an impact on the travel time are investigated. An obvious candidate is the travel distance. The relationship between travel time and estimated travel distance will be discussed in more detail in the next section, and so only a brief mention will be made here. Figure 4-6 shows how the travel time varies by (rectilinear) distance for the city of Calgary using cleaned data (dataset 4) and considering only the highest priority calls. The relationship does appear to follow the KWH piecewise square root-linear relationship. This will become

clearer in the later section when more focus on the shorter distances is provided to show the "elbow" in the relationship. For now, it is sufficient to say that there is a positive relationship between travel time and distance, as would be expected. Also, note from the figure that the variability in the travel time seems to increase slightly as the distance increases. This is discussed further in the following section.

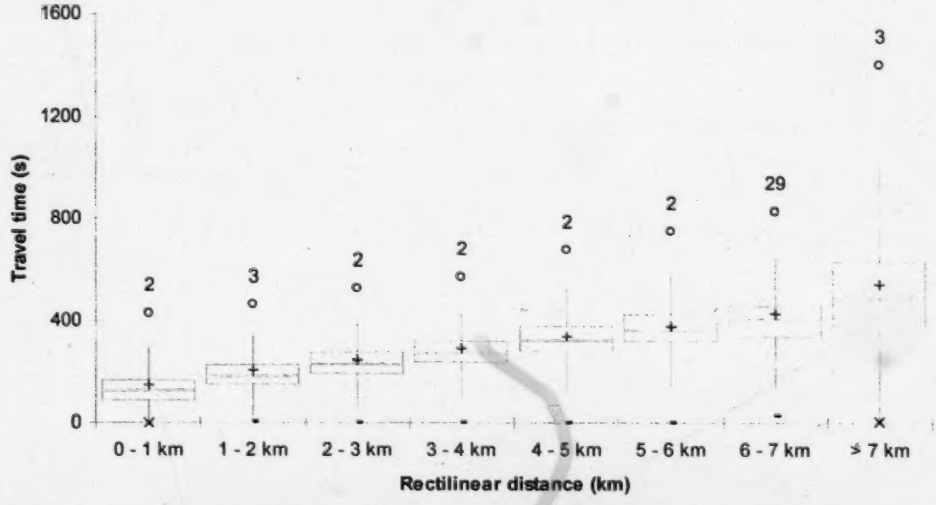


Figure 4 - 6: Box plots for the travel time as a function of distance (observations were grouped into 1 km distance bins) for cleaned data (set 4).

Additional factors that could impact the travel time, and that could be accounted for in a regression model to estimate the travel times, include the type or priority of the call, and seasonal effects such as time of day, day of week, month or season. Because of the strong relationship between travel time and estimated travel distance, it is important to take that into effect when looking at the relationship between travel time and other variables. One way to do this is to examine a specific pair of locations so that the distance will not be a major factor. Figure 4-7 shows the distribution of travel times for a particular demand location (uncleaned data). The average travel time for ambulances responding to calls at this location is just over five minutes with a standard deviation of almost three

and a half minutes. However, as indicated in the histogram, the travel times are skewed to the right, with a few very long times. If we examine the cases with the highest travel times, it is apparent that they may be due to errors in the data. The seven records with the highest travel times have on-scene times that are unexpectedly short (five of them are under 3 minutes, which is in the bottom two percentile). Additionally, the record with the lowest travel time has an unexpectedly long on scene time (in the top two percentile). If we remove these records, the mean and standard deviation drop substantially (by just over 20 seconds and about 70 seconds respectively). Although these records would be flagged as potential outliers if the data were transformed (using a log transformation) to provide a more symmetric distribution, they would not be indicated as outliers in the untransformed data because of the heavy skewness of the data. For this reason, as mentioned in a previous section, it makes sense to use more robust statistics (than the mean and standard deviation) to describe the distribution.

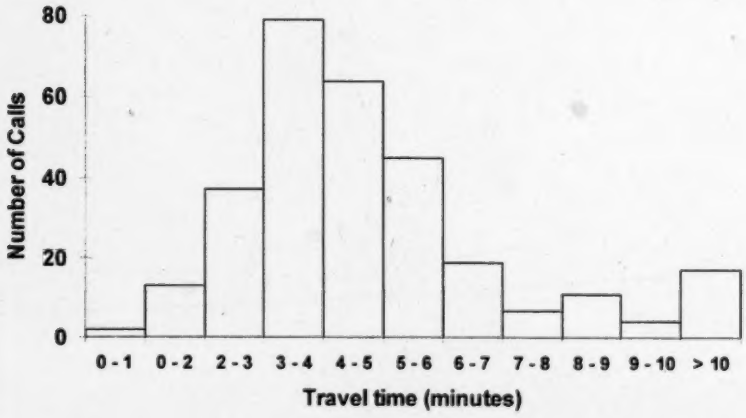


Figure 4 - 7: Travel time distribution for a particular demand location.

Figure 4-8 shows the CDF and box plot summaries for vehicles dispatched from the closest station versus those dispatched from all other locations (including other stations, the road, and alternate locations such as hospitals). The distance

between this demand location and the closest station is about 2.4 km. The median travel time for all responses and for those from the closest station only is almost four and a half minutes and just over four minutes respectively, while the middle half of the data has a range of about two and a half minutes and just under two minutes for these two cases. Note that this demand location was centrally located and in the middle of about three fairly close stations and thus the difference between all responses and "closest station only" responses is not as great as would be expected for more peripheral demand locations.

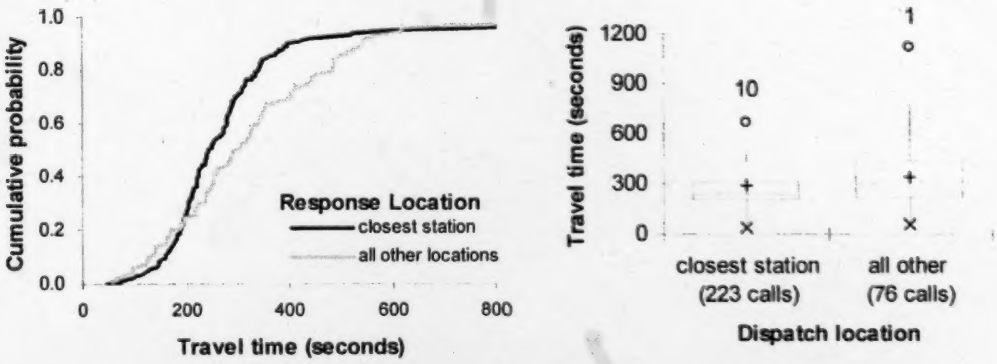


Figure 4 - 8: Cumulative distribution functions and summary box plots for the travel times of responses to a particular demand location, from the closest station and from all other locations.

Next, the calls responded to by an ambulance from the closest station were grouped by priority and the corresponding plots are shown in Figure 4-9. Note that the group sizes are fairly small when the calls are divided in this way especially for the higher priority (Charlie and Delta) calls. Although the median travel time decreases as the priority increases as would be expected, the three outliers and increased variability indicated in the box plot for the most serious (Delta) calls warrant further investigation. In looking specifically at the three serious outliers, it is apparent that these responses all occurred during the rush hour times of the day.

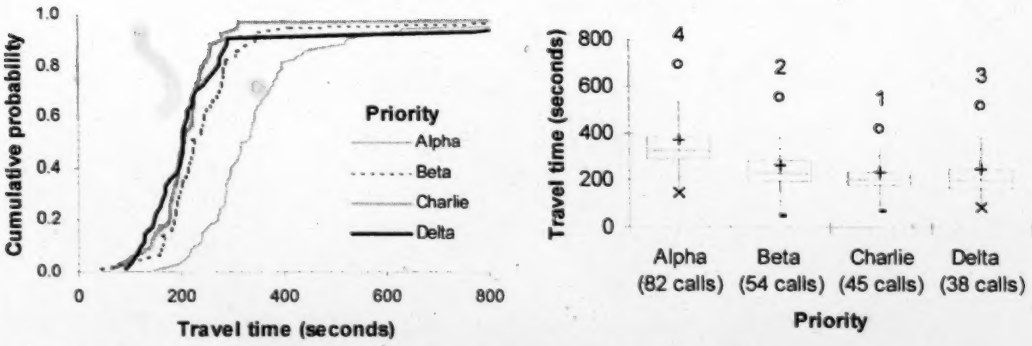


Figure 4 - 9: Cumulative distribution functions and summary box plots for the travel times to a particular demand location from the closest station grouped by priority. Priority increases from left to right in the second frame with Delta calls representing the most serious.

In order to examine the effect of rush hour traffic on the travel times, the calls for the same station demand node pair were grouped according to whether or not the response occurred during rush hour times (between 8:00 and 10:00am or between 3:00 and 5:00pm). The results are shown in Figure 4-10. As indicated in these plots, the responses during rush hour times do tend to be longer and also show greater variability in the travel time. In summary, there are a number of factors that can contribute to the variability of the travel time data in addition to the distance to the closest station (such as the location of responding unit, the priority of the call, and the time of the day that the travel takes place).

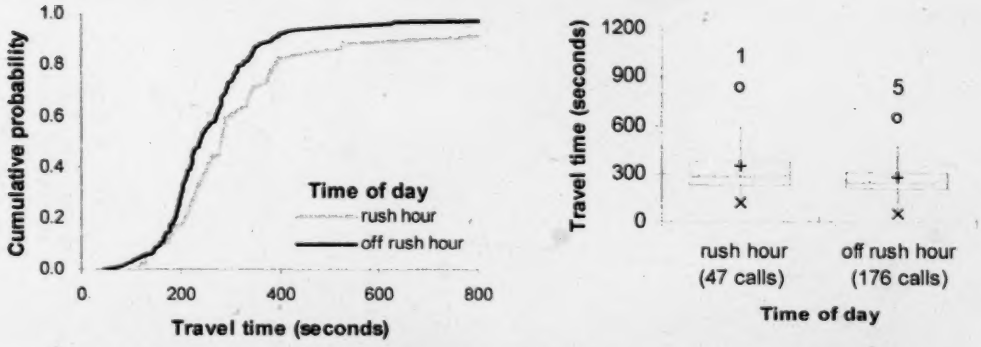


Figure 4 - 10: Cumulative distribution functions and summary box plots for the travel times of responses to a particular demand location, from the closest station grouped into rush hour and off rush hour responses.

Next we examine some travel time distributions to see if they can be modelled using a particular theoretical distribution. A distribution that we are specifically interested in testing the travel times against is the lognormal distribution. This distribution is described as playing a fundamental role in the physical, biological and social sciences, in a paper by Limpert, Stahel, and Abbt, (2001). The authors discuss the link between the lognormal distribution and multiplicative variability and provide other insights into the role of the lognormal distribution. In terms of the variability in ambulance travels times, it would make sense for the lognormal distribution to provide a good model since the travel time is the average travel speed multiplied by the distance and many of the things that can have an impact on the travel times would do so by affecting the travel speed. Advantages of using the lognormal distribution are that it restricts travel times to be positive, the distribution can be skewed, and calculations are no more difficult than with a normal distribution.

Figure 4-11 shows the CDFs of the travel times for a particular demand location. This is the same demand location that we discussed in detail above, and the eight outliers correspond to those mentioned as having extreme values in the time spent on-scene (the records with the smallest and seven largest travel times). As can be seen in the CDFs, after removing the outliers, the empirical distribution seems to be quite close to the lognormal distribution. A Q-Q plot of the sample quantiles (for the logarithm of the travel times) vs. the normal quantiles for the full data and the data with the eight outliers removed is shown in Figure 4-12. A test for lognormality based on the correlations between these quantiles (Johnson and Wichern, 1982) indicates that after removal of the eight outliers, we do not reject the hypothesis of normality for the logarithms of the travel times (and hence for lognormality of the travel times) at a 1% level of significance.

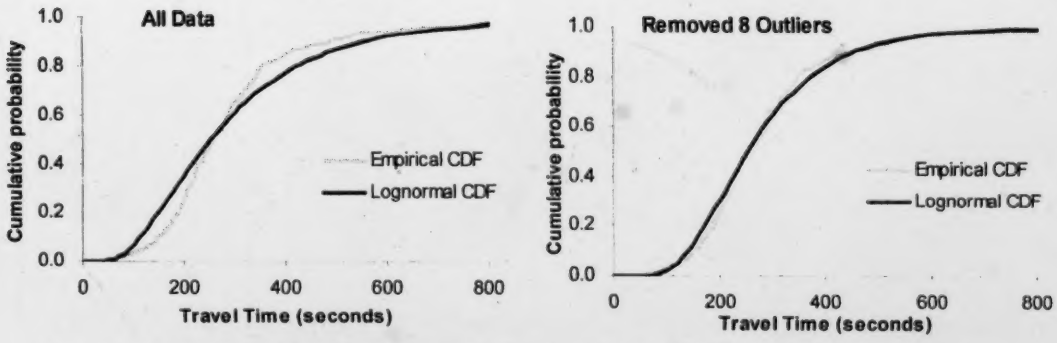


Figure 4 - 11: Empirical and theoretical cumulative distribution functions before and after the removal of outliers.

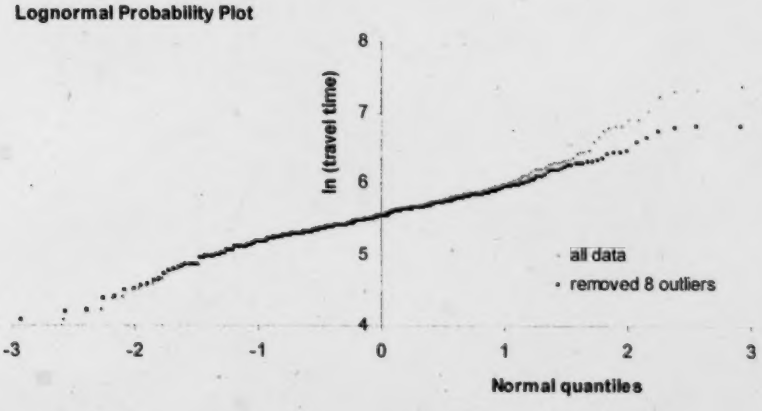


Figure 4 - 12: Q-Q plot of the log of the ordered travel times vs. the corresponding normal quantiles.

Another example that includes all of the responses to a particular area (as indicated by the intersection) provides similar results (see Figures 4-13, and 4-14). In this case, the data seem to be close to lognormally distributed with the exception of a small proportion of the data points with very small travel times. If the records with travel times lower than 20 seconds are removed, then the hypothesis test for lognormality of the travel times is not rejected (at a 5% level of significance). Note that the median and interquartile range for this data are 168 s and 166 s respectively, and the travel times below 20 s represented about 3% of the sample.

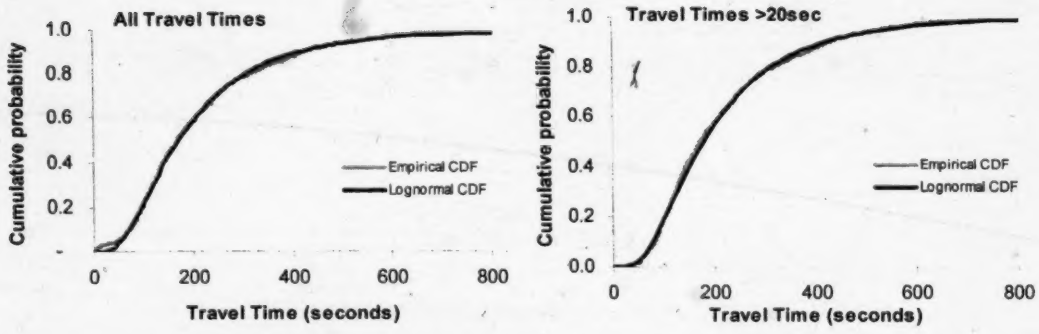


Figure 4 - 13: Empirical and theoretical cumulative distribution functions before and after the removal of outliers.

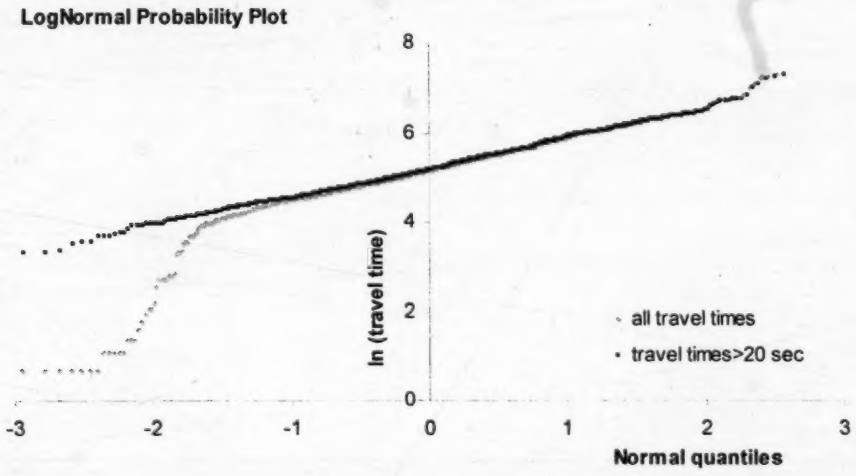


Figure 4 - 14: Q-Q plot of the log of the ordered travel times vs. the corresponding normal quantiles.

It is possible that the travel time distributions observed from the data are actually mixture distributions. Considering the Q-Q plots just discussed, the lines for the original datasets resemble some of the plots in a paper by Burmaster and Thompson (1999) that combine two lognormal distributions. In particular the line for the original data in Figure 4-13 looks like the results for a mixed distribution where the main distribution is a lognormal distribution with a small proportion

(<5%) of observations from a different lognormal distribution with a lower mean and higher standard deviation. As mentioned for this data, there are a small number of very low travel times (less than 20 seconds) that do not seem to fit with the remainder of the distribution. Perhaps combinations of true travel time distributions and error distributions due to disruptions in the recording process are what led to the observed distribution.

Modelling Emergency Vehicle Travel Times

As discussed above, there is an obvious relationship between travel time and distance. The box plots shown in Figure 4-6 do not give a clear picture of some of the difficulties in estimating this relationship from actual data, so some example scatterplots are provided here to illustrate the issues. Outliers have already been discussed, but to give an idea of what the data actually look like, we show the same scatterplot twice with a wide scale on the axes and more narrowly focused into the relevant range of the data (Figures 4-15 and 4-16).

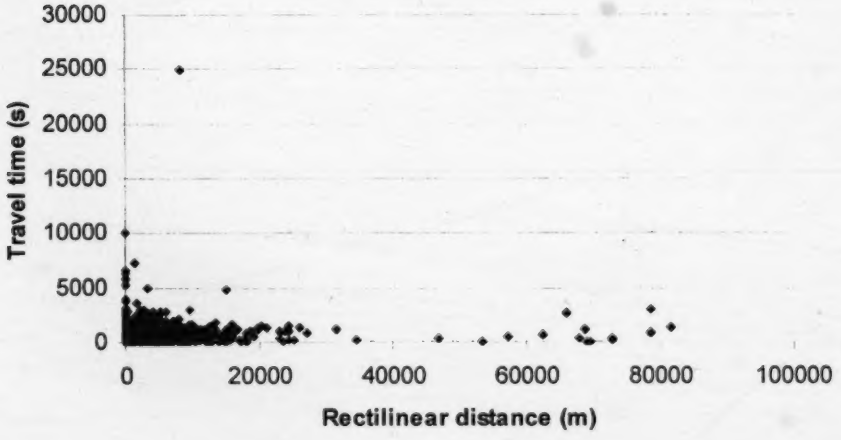


Figure 4 - 15: Scatterplot of travel time vs. distance for original (uncleaned) data.

There are some obvious outliers in the first scatterplot, and outliers such as these can have a dramatic effect on an estimated regression line. Once the focus is on the relevant range a clear pattern is seen, but as indicated in the second scatterplot there is still a lot of variation and even more outliers that are likely the result of error. For example, there are a number of points along the x-axis at various distances that show close to zero travel time. This second scatterplot also shows the estimated relationship, using the method proposed by Kolesar, Walker, and Hausner, (1975) and outlined in Appendix 3.

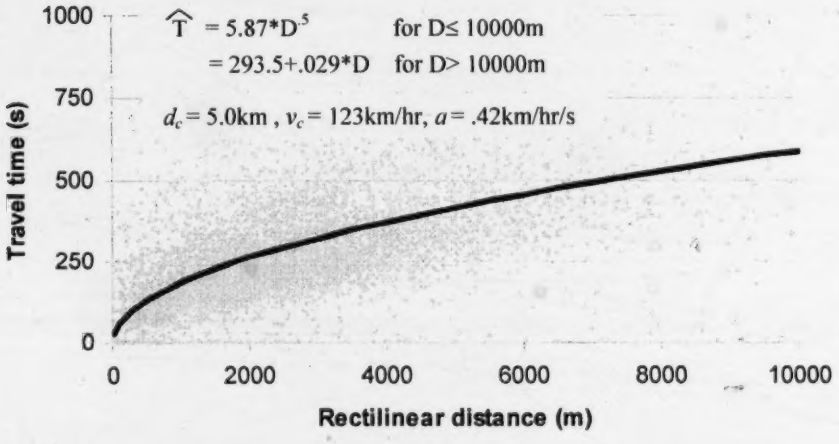


Figure 4 - 16: Scatterplot of travel time vs. distance for original (uncleaned data), focusing on relevant range.

The unlikely parameters (shown in Figure 4-16), obtained using the uncleaned data highlight the dramatic impact that outliers can have on the estimated function. Results closer to those obtained in previous studies were obtained when the outliers were removed before estimating the travel time function. However, there are some other issues (discussed next) that suggest that a more robust method for estimating the function should be considered, so we will not present any further results with this method.

An important note to make here is that the travel distance (the explanatory variable) is subject to error as is the dependent variable. This can complicate the estimation of the relationship between the two. Another difficulty is that the residuals from the regression analysis don't appear to be normally distributed with constant variance about the estimated mean curve as shown in Figure 4-17.

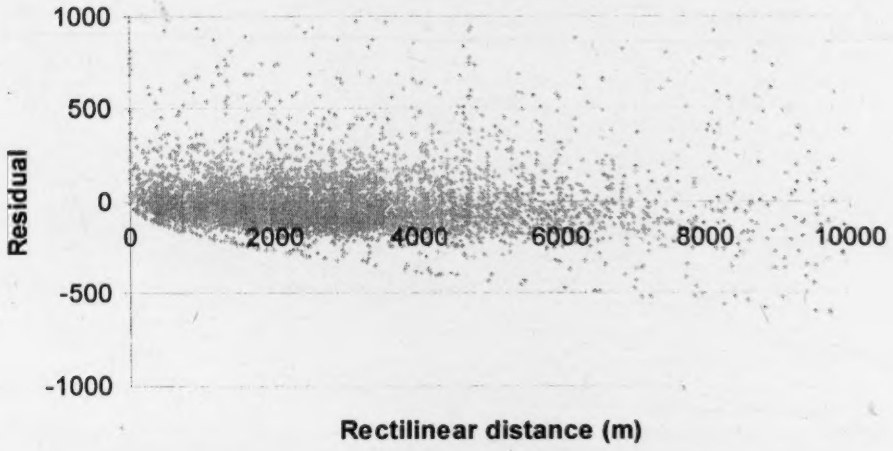


Figure 4 - 17: Scatterplot of the residuals of the estimated KWH travel time function.

One method that may help to alleviate some of these problems is grouping the observations. Thus, the observations were grouped into distance bins of a half a kilometre in width and the relationship between the median distance and travel time for the groups was estimated (again using the method outlined in Appendix 3). As shown in Figure 4-18 the medians follow a very predictable relationship.

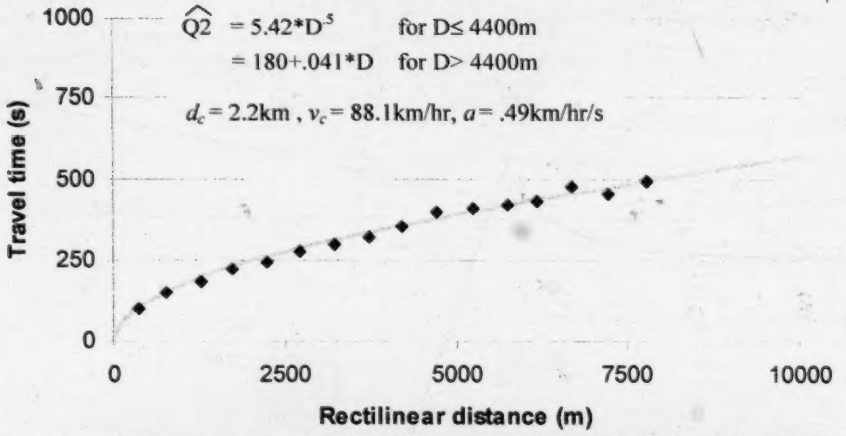


Figure 4 - 18: Scatterplot of the median travel times for each 500 m distance group and the estimated KWH travel time function.

The parameters (shown in Figure 4-18) for the function obtained using the method of grouping the data and estimating the median travel time are more realistic as compared to the results from the ordinary least squares regression given earlier (in Figure 4-16). As expected and as evidenced in the graph this method is more robust to outliers in the data. The parameters here can not be compared directly to those obtained in previous studies since they are for estimating the median (and not the mean) travel time.

Up to this point, the focus of this section has been on the relationship between the level of the travel time distribution and travel distance. Next, the relationship between the spread and the level of the data is considered as well as a possible relationship between the variability in the travel times and the distance variable. For example, is there more variability in the travel time for higher average travel times or for longer travel distances? A plot of the spread vs. the level of the dataset is shown in Figure 4-19. It is apparent that the variability of the travel times is not constant regardless of the level whether considering the standard deviation and the mean, or the interquartile range and the median. The implication is that the variation can not be estimated independently from the level.

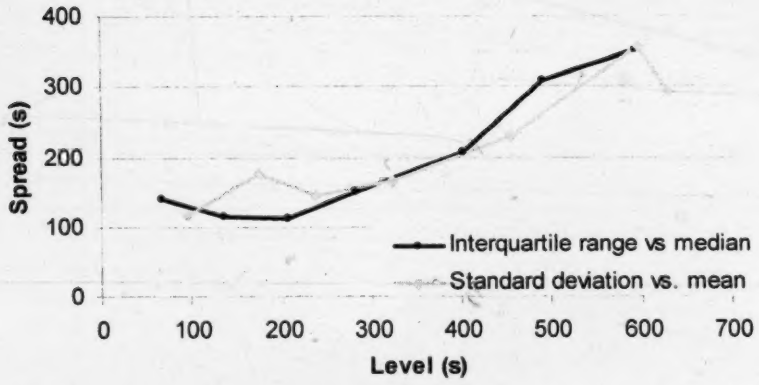


Figure 4 - 19: Spread vs. level plot for the travel time divided into level bins based on estimated travel time.

However, as discussed above, the lognormal distribution seems to provide a good model for ambulance travel times, and a natural measure for the variability of a lognormal distribution is the multiplicative standard deviation (Limpert, Stahel, and Abbt, 2001), σ^*, which is related to the standard deviation of the corresponding normal distribution, σ, by $\sigma^* = e^{\sigma}$. There is a simple relationship between the percentiles of a lognormal distribution and the percentiles of the corresponding normal distribution (Aitchison and Brown, 1957), which leads to a natural robust estimate for σ^*. The estimator is $(Q_1/Q_2)^{-1.483}$, where Q_1 and Q_2 are the first quartile and the second quartile (or median) of the lognormal distribution. Using a grouping method to estimate the first quartiles of the travel time as a function of distance (similar to the method discussed above for estimating the median), it is apparent that the best relationship is not a piecewise square root linear relationship (as in the KWH function), but rather a simple power relationship. Figure 4-20 shows the first and second quartiles of the travel time for the groups, along with the relationships estimated between these and the rectilinear distance.

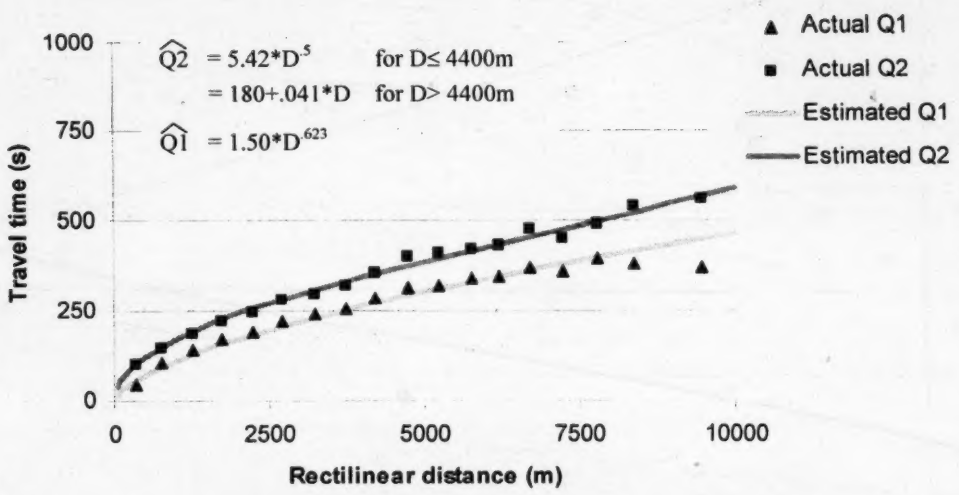


Figure 4 - 20 Scatterplot of the first and second quartiles of the travel times for each 500 m distance group and the estimated relationships with the rectilinear distance.

Using these estimated relationships, an estimate for the multiplicative standard deviation of the travel time distribution (depicted in Figure 4-21) was obtained.

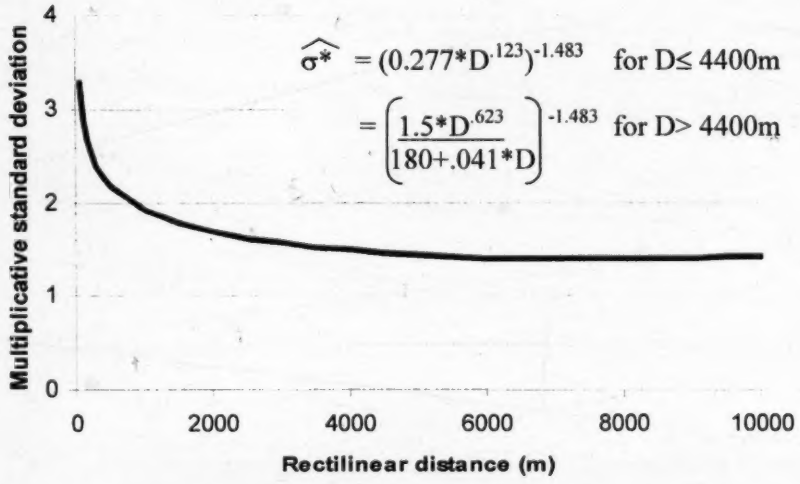


Figure 4 - 21 Estimated relationship of the multiplicative standard deviation of the travel times with the rectilinear distance.

For example, the estimated relationship predicts that at a distance of 2 km, about 68% of the travel times are between 2.4 minutes and 6.8 minutes, while for 5 km, about 68% of the travel times are between 4.5 minutes and 9.1 minutes. As is evident in the figure, the multiplicative standard deviation is relatively stable for most of the relevant range, with the exception of the very low travel distances.

Conclusions and Further Research

This chapter has given a broad overview of the problem of estimating travel times for use in models of emergency service operations. This research may also be more broadly applicable for any service organization where travel is an important factor in the provision of service, for example in pick-up and or delivery services, transporting services, or emergency repair services.

We have been very fortunate to have access to such a wealth of data including the event data that have been the foundation of most of the analysis in this chapter as well as the AVL data, which gave us a unique opportunity to obtain insight into the actual movements of ambulances as they responded to calls. The major theme from the data, and as a result a particular focus of this chapter, is the tremendous variation inherent in these travel times. Although some of the variation is likely due to errors in the data, there are significant patterns in the data that can be used to provide an estimate of the travel time distribution for a particular response to service.

There are many opportunities for future research in this area. The first is in the area of advancing the state of regression-type estimation models specific to this problem. As discussed there are a number of complications in particular with the time-distance relationship, including data issues (multiple outliers often with significant leverage to affect the estimated relationship), non-linearity, non-constant parameters, heterogeneity of variance, non-normal distributions (of

residuals) and substantial variation, some of which may be explainable by other factors which could have an effect on the nature of the estimated relationship. Given the results that we have shown, a useful extension might be to further explore using the KWH function (or some other non-linear relationship) in concert with network distances on various road types (as opposed to using distance predicting functions) and incorporating additional explanatory variables. Another line of research that could have an impact in this area is investigation of more complex functions, with acceleration and deceleration combined with constant speed travel that would allow a (possibly random) number of such components of varying length. Next, in contrast to all other data sources that we know of, AVL data allow one to reconstruct actual routes, rather than just the locations of the origin and the destination. Thus, further research using AVL data could provide more insight into the travel of emergency service vehicles. Additionally, more study into the nature of the distribution of the travel times would be beneficial. Specifically, examination of mixed lognormal distributions could be a fruitful area for investigation. Finally, another important extension would be to do some sensitivity analysis to see what impact various assumptions about the travel time distribution would have on the estimated performance of the system. For example, how would the estimated coverage of a particular system configuration change under different assumptions about the form and parameters of the travel time distribution? Such sensitivity analysis could be helpful in determining which aspects of the distance – travel time relationship could benefit most from greater accuracy.

References

J. Aitchison, and J. A. C. Brown (1957). *The Log-normal Distribution.* Cambridge University Press, Cambridge, UK.

A. A. Aly and J. A. White (1978). Probabilistic Formulation of the Emergency Service Location Problem. *Journal of the Operational Research Society* **29** 1167-1179.

M. L. Brandeau and R. C. Larson (1986). Extending and Applying the Hypercube Queueing Model to Deploy Ambulances in Boston. *TIMS Studies in the Management Sciences* **22** 121-153.

J. Brimberg and R. F. Love (1992). A New Distance Function for Modeling Travel Distances in a Transportation Network. *Transportation Science* **26** 129-137.

D. E. Burmaster, and K. M. Thompson (1999). Using Animated Probability Plots to Explore the Suitability of Mixture Models With Two Component Distributions. *Risk Analysis* **19** 1185-1192.

T. M. Cook, and R. A. Russell (1980). Estimating Urban Travel Times: A Comparative Study. *Transportation Research A* **14A** 173-175.

Y. M. Carson and R. Batta (1990). Locating an Ambulance on the Amherst Campus of the State University of New York at Buffalo. *INTERFACES* **20** 43-49.

K. Chelst and J. P. Jarvis (1979): Estimating the Probability Distribution of Travel Times for Urban Emergency Service Systems. *Operations Research* **27** 199-204.

M. S. Daskin (1987): Location, Dispatching, and Routing Models for Emergency Services with Stochastic Travel Times. *Spatial Analysis and Location-Allocation Models*, A. Ghosh and G. Rushton, eds., Van Nostrand Reinhold Co., New York, 224-265.

M. S. Daskin, and A. Haghani (1984). Multiple Vehicle Routing and Dispatching to an Emergency Scene. *Environment and Planning A* **16** 1349-1359.

E. Erkut, R. Fenske, S. Kabanuk, Q. Gardiner, and J. Davis (2001). Improving the Emergency Service Delivery in St. Albert. *INFOR* **39** 416-433.

Geomatics Canada (1993). *GPS Positioning Guide*, Geomatics Canada Geodetic Survey Division, Ottawa, Canada.

J. Goldberg, R. Dietrich, J. M. Chen, M. G. Mitwasi, T. Valenzuela, and E. Criss (1990). Validating and Applying a Model for Locating Emergency Medical Vehicles in Tucson, AZ. *European Journal of Operational Research* **49** 308-324.

J. Goldberg and L. Paz (1991). Locating Emergency Vehicle Bases when Service Time Depends on Call Location. *Transportation Science* 25 264-280.

S. G. Henderson, and A. J. Mason (2000). Development of a Simulation and Data Visualisation Tool to Assist in Strategic Operations Management in Emergency Services. School of Engineering Technical Report 595, University of Auckland, New Zealand.

S. G. Henderson, and A. J. Mason (2004). Ambulance Service Planning: Simulation and Data Visualisation. To appear in *Handbook of OR/MS Applications in Healthcare*, eds. F. Sainfort, M. Brandeau, and W. Pierskalla, Kluwer.

R. A. Johnson, and D. W. Wichern (1982). *Applied Multivariate Statistical Analysis*. Prentice Hall, Englewood Cliffs, NJ.

P. Kolesar, and E. H. Blum (1973). Square Root Laws for Fire Engine Response Distances. *Management Science* 19 1368-1378.

P. Kolesar (1975). A Model for Predicting Average Fire Engine Travel Times. *Operations Research* **23** 603-613.

P. Kolesar, W. Walker, and J. Hausner (1975). Determining the Relation between Fire Engine Travel Times and Travel Distances in New York City. *Operations Research* **23** 614-627.

R.C. Larson (1974). A Hypercube Queueing Model for Facility Location and Redistricting in Urban Emergency Services. *Computers and Operations Research* **1** 67-95.

R.C. Larson (1975). Approximating the Performance of Urban Emergency Service Systems. *Operations Research* **23** 845-868.

R.C. Larson , and A. R. Odoni (1981). *Urban Operations Research* Prentice Hall, Englewood Cliffs, NJ.

E. Limpert, W. A. Stahel, and M. Abbt (2001). Log-normal Distributions across the Sciences: Keys and Clues. *BioScience* **51** 341-352.

R. F. Love and J. G. Morris (1979). Mathematical Models of Road Travel Distances. *Management Science* **25** 130-139.

P. B. Mirchandani and A. R. Odoni (1979). Locations of Medians on Stochastic Networks. *Transportation Science* **13** 85-97.

G. Strang and K. Borre (1997). *Linear Algebra, Geodesy, and GPS.* Wellesley-Cambridge Press, Wellesley, MA.

H. Uster and R. F. Love (2003). Formulation of Confidence Intervals for Estimated Actual Distances. *European Journal of Operational Research* **151** 586-601.

W. E. Walker, J. M. Chaiken, and E. J. Ignall, editors (1979). *Fire Department Deployment Analysis: a Public Policy Analysis Case Study / The Rand Fire Project.*

Chapter 5: Conclusions and Further Research

Summary

The topic of this dissertation is uncertainty in the response time of ambulances and the resulting impact on the performance of an EMS system. Three major factors influencing the uncertainty in the response time have been considered in detail with a chapter devoted to each one. The second chapter focuses on the effect that uncertainty in ambulance availability has on the response time and estimated coverage of the system. In that chapter, a procedure to estimate the ambulance utilization and dispatch probabilities is developed that relaxes certain operational assumptions made in previous methods. The focus of the third chapter is on the second component, pre-travel delay, and how it can impact the measured performance of the system. In that chapter, an optimal location model that incorporates all three of the components at the heart of this dissertation is detailed. Although all of the components are incorporated into the model formulation, the focus is on the pre-travel delay, and results for that component are provided. The fourth chapter focuses on the third component, the travel time to the scene of the emergency. The estimation of this component is complicated by its relationship with travel distance, which is itself a difficult factor to estimate. Each of these three factors brings its own challenges, and each impacts the uncertainty in the response time in a different way. Together the three can have a major impact on the estimated performance of an EMS system as well as on prescriptions for optimal system design.

Conclusions

The principal finding is that the inclusion of uncertainty in response time can have a considerable impact on the estimated performance of an EMS system. In the second chapter, details were given to show that the coverage of the system as a whole, as well as of particular regions, can vary substantially depending on how uncertainty in ambulance availability is modeled. Chapter 3 gave similar results for the pre-travel delay component of the response time, and further indicated that modelling the uncertainty in this component can impact the solutions provided by a prescriptive model to locate ambulances. Finally, although the fourth chapter did not provide calculations of the impact of the uncertainty in travel time on the measured performance, this component enters into the location model of Chapter 3 in the same way as the pre-travel delay component and can be expected to impact the estimated coverage in a similar fashion. Chapter 4 did indicate that there is a great degree of variability in travel times and that the distributions are positively skewed (lognormal distributions provided a good fit in most cases), which both would be expected to increase the impact on the estimated coverage in a realistic setting. Further, since estimates accurate to within a small degree of error can be necessary for planning decisions in such systems, it is desirable to use models that make as few simplifying assumptions as possible. Thus, the topics presented in this dissertation represent a step towards merging the descriptive and prescriptive literature on EMS operations.

Further Research

Research that can be undertaken to extend each of the three topics detailed in this dissertation is discussed in the corresponding chapters. In general, a fruitful area for future research would be to extend the methods in this dissertation, or to develop additional methods, that can enhance the realism of prescriptive models for EMS operations. For example, one criticism of such models is that they do

not consider responses from locations other than a station or some other designated location (i.e., they don't allow for responses from the road). Methods that estimate travel time could be extended to incorporate this type of response.

Another important operational characteristic of an EMS system that has been difficult to deal with, in terms of its impact on prescriptive models for tactical decisions, is ambulance redeployment. Although there have been a number of contributions in terms of dynamic redeployment at the operational level, (for example Gendreau, Laporte, and Semet, 2001; Kolesar and Walker, 1973), a need has been identified for methods that deal with redeployment in models at the tactical level. Redeployment is an operational issue, and research that deals with it as such is valuable. However, it is also important to consider redeployment at the tactical level since the redeployment policy used can have a large impact on the effectiveness of the system. If models used at the tactical level ignore the redeployment policy used, then they could lead to bad tactical decisions. The open research question is how to incorporate redeployment in models at the tactical level, in a way that is tractable and leads to improved tactical decisions.

Despite the fact that the focus in this dissertation has been on a particular measure of system performance, the system coverage, it is recognised that there are many other important measures of performance for EMS systems. Some of these, such as measures of workload imbalance, average travel time, and the probability that all servers are busy, have been discussed and are easy to incorporate into the models developed here. However, ideally, one would want to consider direct outcome measures, such as measures of survival and reductions in disability and suffering, when making operational decisions for EMS systems. While there has been a great deal of literature on the relationship between response time and survival in the case of victims of cardiac arrest (Cretin and Willemain, 1979; Eisenberg, Bergner, and Hallstrom, 1979; Stiell et al., 1999), there is less literature on the measures of disability in such cases, and little literature on these

relationships for other types of emergencies. This type of research can be quite challenging especially given the many potential, and often difficult to measure or estimate, intervening variables such as the time interval between when the symptoms begin, until the call is made to request an ambulance response. However, it has the potential to dramatically improve the worth of operational research methods for planning decisions in EMS systems.

References

S. Cretin, and T.R. Willemain (1979). A Model of Prehospital Death from Ventricular Fibrillation following Myocardial Infarction. *Health Services Research* **14** 221-234.

M. Eisenberg, L. Bergner, A. Hallstrom (1979). Paramedic Programs and Out-of-Hospital Cardiac Arrest: I. Factors Associated with Successful Resuscitation. *American Journal of Public Health* **69** 30-38.

M. Gendreau, G. Laporte, F. Semet, (2001). A Dynamic Model and Parallel Tabu Search Heuristic for Real-time Ambulance Relocation. *Parallel Computing* **27** 1641-1653.

P. Kolesar, and W. E. Walker, (1973). An Algorithm for the Dynamic Redeployment of Fire Companies. *Operations Research* **22** 249-274.

I. G. Stiell, G. A. Wells, V. J. DeMaio, D. W. Spaite, B. J. Field, D. P. Munkley, M. B. Lyver, L. G. Luinstra, R. Ward, (1999). Modifiable Factors Associated with Improved Cardiac Arrest Survival in a Multicenter Basic Life Support/Defibrillation System. *Annals of Emergency Medicine* **33** 44-50.

Appendices

Appendix 1: Estimating the Average Busy Fraction

The average fraction of time that an ambulance is busy (not available to respond to calls) is $\lambda\tau/s$, i.e., the average server utilization for an s-server queueing system, assuming that the number of calls "lost" due to queueing is negligible. The average "service time", τ, (during which an ambulance is tied up with a call) can be broken down into the following components: average travel time to the call, average on-scene time, and average time spent traveling to and remaining at a hospital, denoted $\mathrm{E}[T_{\text{to call}}]$, $\mathrm{E}[T_{\text{on scene}}]$, and $\mathrm{E}[T_{\text{hospital}}]$, respectively.

Consequently, the average busy fraction can be expressed as $\lambda(\mathrm{E}[T_{\text{to call}}]+\mathrm{E}[T_{\text{on scene}}]+\mathrm{E}[T_{\text{hospital}}])/s$. The arrival rate λ as well as two of the three components of the average service time, the average on-scene time and the average time spent traveling to and being at a hospital, are exogenous input. The average travel time to a call can be expressed as $\mathrm{E}[T_{\text{to call}}]=\sum_{m=1}^{M} h_m \sum_{j=1}^{J} f_{jm}(\mathbf{x})\,\mathrm{E}[T_{jm}]$. This leads to the following formula for approximating ρ as a function of $\mathbf{x}$:

$$\rho(\mathbf{x})=\frac{\lambda}{s(\mathbf{x})}\left\{\sum_{m=1}^{M} h_m \sum_{j=1}^{J} f_{jm}(\mathbf{x})\,\mathrm{E}[T_{jm}]+\mathrm{E}[T_{\text{on scene}}]+\mathrm{E}[T_{\text{hospital}}]\right\} \qquad (12)$$

The derivation of this formula required some approximations. In particular, the time spent traveling back to a station from the hospital is excluded from the average service time since the ambulance is available to respond to incoming calls during this time. On the other hand, the expression for $\mathrm{E}[T_{\text{to call}}]$ assumes that all calls are responded to from an ambulance at a station.

Appendix 2: Derivation for the KWH Function

In this appendix a derivation of the KWH travel time function is provided. Let D be the distance of the trip and let T be the travel time. Let t_c be the time, and d_c the distance, needed to reach the cruising speed, i.e., $t_c = v_c / a$. Figure 4-22 shows graphs of the acceleration, velocity, and distance as functions of time for the case where the trip is long enough that the cruising speed is reached, i.e., $T > 2t_c$.

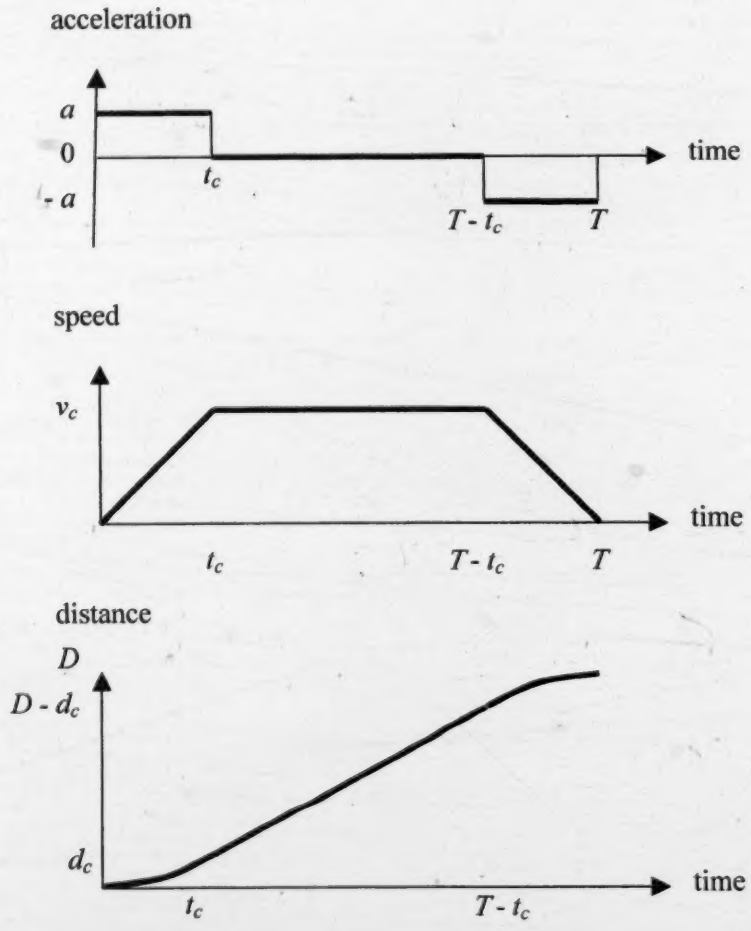


Figure 4 - 22 Pictures of the acceleration, velocity and distance as functions of time for the KWH travel time relation.

Let $v(t)$ be the speed at time t and $d(t)$ the distance traveled by time t. From the graphs, we find that

$$v(t) = at \text{ for } t \le t_c \Rightarrow d_c = d(t_c) = \int_0^{t_c} v(t)dt = a\int_0^{t_c} t dt = \tfrac{1}{2}at_c^{\,2} = \tfrac{1}{2}a\left(\frac{v_c}{a}\right)^2 = \frac{v_c^2}{2a}$$

Also, we have

$$D = \int_0^T v(t)dt = \int_0^{t_c} v(t)dt + \int_{t_c}^{T-t_c} v(t)dt + \int_{T-t_c}^{T} v(t)dt$$

$$= d_c + v_c \int_{t_c}^{T-t_c} dt + d_c = 2d_c + v_c(T - 2t_c) = v_c T - \frac{v_c^2}{a}$$

$$\Rightarrow T = \frac{D}{v_c} + \frac{v_c}{a}$$

When the trip is so short that cruising speed is not reached, we have

$$D = \int_0^T v(t)dt = \int_0^{\frac{1}{2}T} v(t)dt + \int_{\frac{1}{2}T}^{T} v(t)dt = 2\int_0^{\frac{1}{2}T} v(t)dt = 2a\int_0^{\frac{1}{2}T} dt = \frac{aT^2}{4}$$

$$\Rightarrow T = 2\sqrt{D/a}$$

Combining these results, we get

$$T = \begin{cases} 2\sqrt{D/a} & \text{for } D \le 2d_c \\ D/v_c + v_c/a & \text{for } D > 2d_c \end{cases}$$

Appendix 3: Method for Fitting the KWH Function

In order to fit the piecewise square root-linear relationship to travel time and distance data, Kolesar, Walker, and Hausner, (1975) proposed a search method where one parameter is varied and the (least squares) best fit for the other parameter(s) are calculated based on that parameter. In their formulation, they performed a weighted least squares regression with the number of observations at each distance used as weights. Since the data that they used were based on field experiments and the distances were obtained from odometer readings measured to tenths of a mile, it was natural to use the weighted least squares formulation. However, for the purposes of the analyses in this chapter, since distances are calculated using distance metrics and the data are from very large databases, this method is not natural and so the formulation given below has been modified slightly from that given in the paper by Kolesar, Walker, and Hausner.

To simplify the notation, we will use the following for the travel time function.

$$f(D_i) = \begin{cases} c\sqrt{D_i} & D_i \le d \\ a + bD_i & D_i > d \end{cases}$$

Then for a set of observations, $i = 1, 2, \ldots N$, the least squares method involves finding the parameters a, b, c, and d, that minimize the sum of squared deviations between the actual travel times, T_i, and those predicted based on the distance, D_i, using this travel time function. For the piecewise function, two further constraints are that the functions meet and have the same slope at $D_i = d$.

$$\text{minimize} \quad \sum_{i=1}^{N}\left(T_i - f(D_i)\right)^2$$

subject to

$$a + bd = c\sqrt{d}, \quad b = c/\left(2\sqrt{d}\right)$$

The parameters a and c can be eliminated by solving for them in terms of b and d, and then the problem is as follows:

$$\text{minimize} \quad \sum_{i=1}^{N_d}\left(T_i - 2b\sqrt{dD_i}\right)^2 + \sum_{i=N_d+1}^{N}\left(T_i - bd - bD_i\right)^2 \qquad (1)$$

where the observations have been ordered in increasing value of D_i and N_d is the largest value of i for which $D_i \le d$. Then, fixing d, the optimal value of b given d, $b^*(d)$, is found by setting the derivative with respect to b of (1) to zero.

$$b^*(d) = \frac{2\sqrt{d}\sum_{i=1}^{N_d}\left(T_i\sqrt{D_i}\right) + \sum_{i=N_d+1}^{N} T_i\left(d + D_i\right)}{4d\sum_{i=1}^{N_d} D_i + \sum_{i=N_d+1}^{N}\left(d + D_i\right)^2}$$

Finally, d can be varied, $b^*(d)$ calculated, and the objective function measured in order to obtain the optimal pair of parameters, b^* and d^*.